Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20, 41 36 60

IO 73/1

11. januar 1973

METODEHEFTE NR. 3

Notater om definisjon av aldersbegrepet, utdanningsdata, kontroll-rutiner ved AKU, og design-effekt ved utvalgsundersøkelser

INNHOLD

FORORD

Metodehefter i serien Arbeidsnotater

I tilknytning til mange prosjekter i Statistisk Sentralbyrå utarbeides det mindre, upretensiøse notater for avklaring av spørsmål av metodisk interesse. Det kan dreie seg om utvalgsteknikk, alternative spørsmålsformuleringer, presentasjonsmetoder, begrepsavklaringer, diskusjon av "funn" i data, systemidéer eller andre temaer. Selv om mange slike notater bare har begrenset interesse i ettertid, vil det blant dem være noen som kunne fortjene å bli mer alminnelig tilgjengelig enn de har vært hittil. Det kan også være nyttig å ha dem registrert sentralt slik at det blir lettere å få oversikt over det stoffet som foreligger, og å referere tilbake til det.

Byrået har innført en publiseringsordning for stoff av dette slaget. En publiserer leilighetsvis et passende antall slike notater samlet i metodehefter i serien Arbeidsnotater. Inneværende hefte er det tredje av denne typen.

Forsker Jan M. Hoem er oppnevnt som redaktør av metodeheftene. Medarbeidere i Byrået som lager stoff som kan være aktuelt, bes sende dette til redaksjonen etter hvert som det blir ferdig.

Kontorlederne bes holde øynene åpne for denne nye publiseringsmuligheten.

Konsulent Bjørn L. Tønnesen er redaksjonssekretær.

Alder - definisjon og klasseinndeling

Av Idar Møglestue

Innledning

Definisjonen av kjennemerket alder varierer i dag fra statistikk til statistikk. Vi nytter således tre ulike grupper av definisjoner, nemlig:

1) alder ved tellingstidspunktet
2) alder ved tidspunktet for visse hendinger
3) alder ved årets begynnelse eller slutt.

I de enkelte statistikker er valget av aldersdefinisjon som hovedregel avgjort ut fra særkrav som stilles innen mer eller mindre snevre fagområder. Det er selvsagt viktig at statistikken så langt råd er søker å imøtekomme slike spesielle behov. Spesialiseringen må imidlertid ikke skje på bekostning av de krav som de mer sammenfattende analyser og de samfunnsmessige planleggingsoppgaver stiller til statistikken.

I et notat av 18. mai 1967 gav Direktøren uttrykk for at det var ønskelig å få en bedre samordning av statistikkene med hensyn på aldersdefinisjon og aldersgrupperinger. Byråets ledelse ville ta dette spørsmål opp til nærmere overveielse[1]. I den anledning ble kontorene bedt om å gi visse opplysninger om praksis når det gjelder definisjon av og gruppering etter kjennetegnet alder.

Aldersbegrepet i statistikken

Kontorene har gitt opplysninger om til sammen 30 enkeltstatistikker med oppgaver over personers alder. Disse statistikker fordeler seg slik etter statistikkområde og aldersdefinisjon:

	I alt	Alder ved		
		tellings-tidspunkt (1)	tidspunkt for visse hendinger (2)	årets beg. eller slutt (3)
Befolkning og helseforhold	8	1	6	1
Jordbruk	1	1	-	-
Lønninger	8	-	-	8
Sosiale forhold	4	-	-	4
Rettsforhold	4	-	3	1
Undervisning	5	-	-	5
I alt	30	2	9	19

1) Retningslinjer for definisjon og klasseinndeling av kjennemerket alder er tatt med som vedlegg til notatet.

I to tredjeparter av statistikkene er alderen definert som fylte år pr. statistikkårets begynnelse eller slutt. Definisjon 3 er således alt nyttet i storparten av statistikkene. Det er i første rekke i befolknings- og helsestatistikkene at praksis er annerledes, her nyttes som hovedregel den beregnede alder i fylte år pr. tellingstidspunktet eller pr. tidspunkt for observasjon av visse hendinger (fødsel, død, flytting, giftermål o.s.v.).

Av de 11 statistikker som baserer seg på definisjon 1 eller 2, mener kontorene at en overgang til bruk av alternativ 3 som hoveddefinisjon, kan anbefales for 4 statistikker. De resterende 7 statistikker tilhører alle området befolknings- og helseforhold. 1. kontor syns ikke å være særlig stemt for overgang til alternativ 3 som hoveddefinisjon. Kontoret legger her særlig vekt på at FN's anbefalinger går ut på at alder skal defineres i samsvar med alternativ 1 eller 2. Dessuten peker 1. kontor på kontinuitetsproblemet. Det uttaler: "Det er på det rene at sammenlikninger av aldersstruktur bakover i tiden kan få visse slagsider ved en eventuell omlegging. Problemet er rimeligvis minst når det gjelder folketellingsmaterialet i og med at tellingstidspunktet som regel har vært ved årets utgang eller maksimalt 2 måneder før årsskiftet. Større konsekvenser i denne sammenheng vil forslaget kunne få i samband med den løpende statistikken (aldersgruppering av døde, ektevigde, skilte o.s.v.)".

Innføring av en felles aldersdefinisjon

Oversikten ovenfor viser at alder defineres forskjellig i statistikkene. Det er også meningsforskjell når det gjelder berettigelsen av en slik forskjellsbehandling. Er det på denne bakgrunn formålstjenlig å innføre en felles hoveddefinisjon av kjennetegnet alder i all offentlig norsk statistikk?

Hovedmålet for utviklingen av norsk personstatistikk er å bygge opp et statistikksystem som - som en overbygning - kan gi grunnlag for oppstilling av et differensiert personregnskap. Dette innebærer at personstatistikkene må samordnes så langt råd er, blant annet gjennom innføring av felles definisjoner og felles grupperingsstandarder m.v. En slik samordning må gjennomføres på en måte som sikrer særinteresser knyttet til enkeltstatistikkene. Særinteresser av mer perifer betydning og særinteresser som det på en rasjonell måte er vanskelig å innpasse i et samordnet opplegg, må vi imidlertid være villige til å ofre på vårt hovedmåls alter.

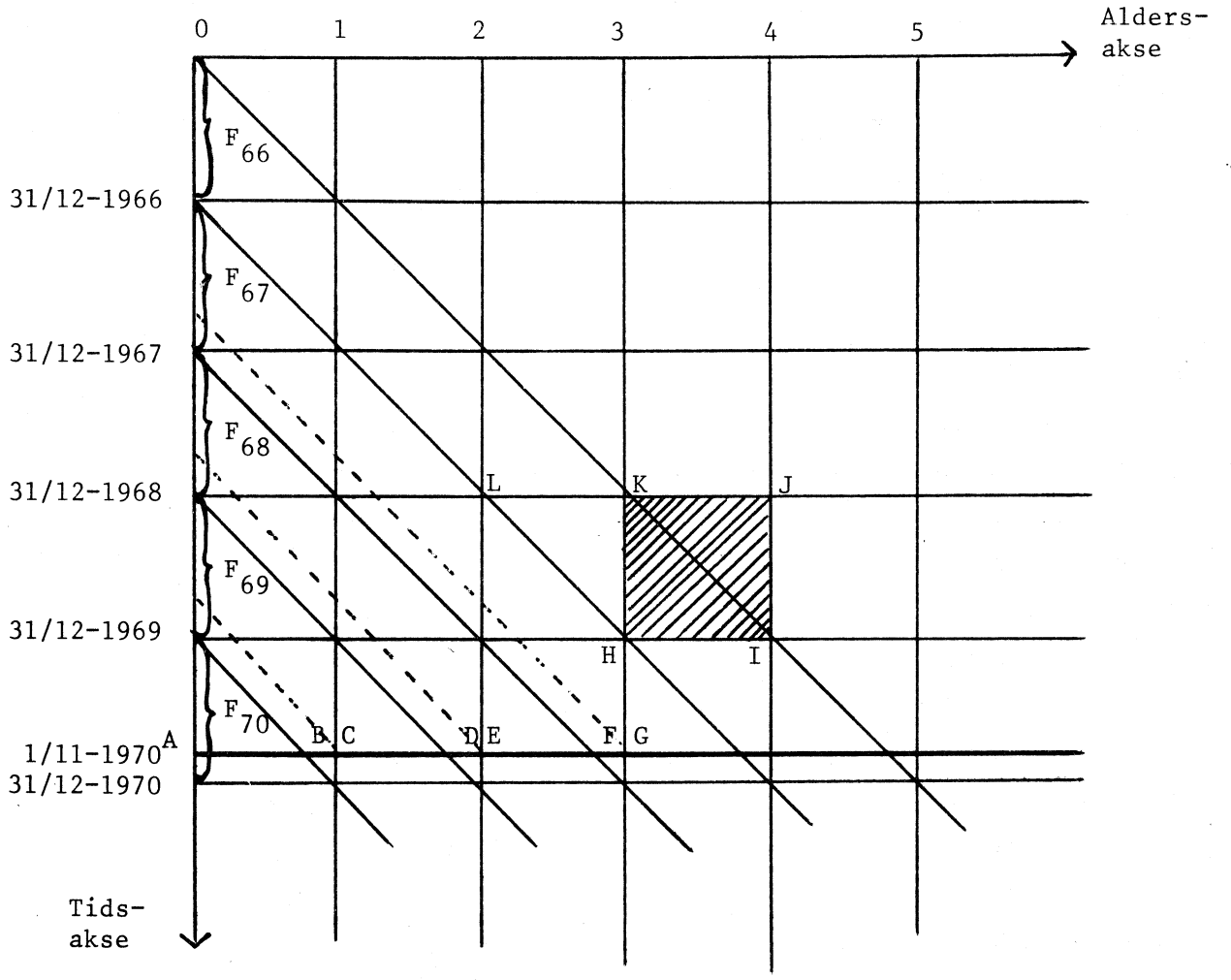
En integrert personstatistikk kan vi skaffe oss på forskjellige måter. En måte er å ordne og lagre grunnmaterialene slik at mulighetene legges best mulig til rette for utkjøring av personstatistiske tabellverk.

Et slikt dataarkivsystem står helt sentralt i planene for en integrert personstatistikk. Planene går dessuten ut på en samordning av de tabellverk som utarbeides og publiseres innen rammen for de enkelte statistikker. Denne samordning har som mål å muliggjøre beskrivelser av persongrupper ved informasjoner hentet fra de ulike tabellverk, noe som forutsetter at enkeltstatistikkene opererer med sammenfallende og entydig avgrensede persongrupper.

Alder er et gjennomgående og viktig kjennetegn i personstatistikken. Ofte er det slik at de persongrupper vi ønsker å analysere nettopp er definert ut fra alderen. Spørsmålet om en felles aldersdefinisjon står derfor helt sentralt i samordningsbestrebelsene. Ja, det forholder seg vel egentlig slik at innføringen av en felles definisjon av alder er helt nødvendig dersom vårt program om en integrert personstatistikk er alvorlig ment.

Av de tre nevnte måter å definere alder på, står alternativ 3 i en særstilling fra et samordningssynspunkt. Etter dette alternativ vil det være en enkel sak å ordne statistikkene slik at disse gir informasjon om de samme persongrupper. Aldersgruppe $a_i - a_j$ i en bestemt årgang t_k av statistikkene vil da på en entydig måte være definert som en delmasse av en bestemt gruppe personer, nemlig som en delmasse av alle personer født i årene $t_{k-i} - t_{k-j}$. En tilsvarende konsistens er det i praksis ikke mulig å oppnå for aldersoppgaver basert på de andre definisjoner.

En eventuell innføring av alternativ 3 som standarddefinisjon av alder vil dessuten by på bearbeidingsmessige fordeler. Alder bestemt på denne måten er nemlig ensbetydende med en gruppering etter fødselsår. Samtidig vil statistikken bli bedre egnet som grunnlag for kohortanalyser. På den annen side er det klart at en slik omlegging vil medføre ulemper. Innen noen statistikker kan, som nevnt foran, sammenliknbarheten med oppgaver for tidligere år bli problematisk. En overgang til beregning av alder pr. årets utgang betyr at noen personer vil bli statistikkført som ett år eldre enn etter tidligere beregningsmåte. Dette vil nærmere bestemt gjelde alle personer med fødselsdato i intervallet mellom tellingstidspunktet (el. tidspunkt for hending) og årets utgang. Tendensen til overvelting til høyere aldersgrupper vil således avhenge av registreringstidspunktets avstand fra statistikkårets utgang. Jo mindre avstand, jo mindre overvelting. Konsekvensene av en endring av aldersdefinisjon er på neste side søkt illustrert grafisk ved hjelp av Lexis' skjema.



I en folketelling for eksempel pr. 1/11 1970 vil aldersgruppen 2 år bestå av alle livslinjer som skjærer linjestykket EG, dersom alderen defineres som fylte år pr. tellingstidspunktet. Defineres derimot alderen som fylte år pr. utgangen av tellingsåret, vil 2-årsgruppen være representert ved linjestykket DF. Den tallmessige konsekvens av definisjonsendringen vil således avhenge av om DE $\gtrless$ FG. Forholdet vil være det samme for alle andre aldersgrupper, bortsett fra 0-årsgruppen. 0-åringene vil etter tradisjonell definisjon utgjøre alle livslinjer som skjærer AC, mens vi ved en ny aldersdefinisjon bare får gruppert AB som 0-åringer. Da BC $\neq$ 0, må vi uten videre konstatere at 0-årsgruppen ikke vil være tallmessig sammenfattende etter de to definisjoner.

Konsekvensene av definisjonsendringen for aldersoppgavene i den løpende årsstatistikken er illustrert ved forskjellen mellom kvadratet HIJK og parallellogrammet HIKL. En registrering av f.eks. dødsfall blant 3-åringer i 1969 vil innebære en opptelling av alle livslinjer som slutter i HIJK. Dersom alderen bestemmes pr. årets utgang, vil det være de avbrutte livslinjer i HIKL som utgjør dødsfallene blant 3-åringer.

Drøftingene ovenfor viser at kontinuiteten er et alvorlig problem ved endring av aldersdefinisjon. For mange statistikker er det neppe tilrådelig å gjennomføre en omlegging uten at kontinuiteten samtidig sikres. Også av hensyn til de problemstillinger som statistikken skal belyse, er det neppe forsvarlig helt å sløyfe beregningene av faktisk alder pr. tidspunktet for vedkommende hending. Eksempelet med registreringen av dødsfall viser blant annet dette. Etter dette synes det som om konklusjonen bør bli denne:

a) Alder ved årets begynnelse eller slutt (alternativ 3) innføres som en gjennomgående aldersdefinisjon i all personstatistikk.
b) Tilleggsoppgaver over alder ved tidspunkt for visse hendinger (alternativ 2) gis i den utstrekning dette anses som nødvendig av hensyn til kontinuiteten eller av hensyn til andre vel begrunnede behov for slike aldersoppgaver.

Klasseinndeling

Statistikkene kan bare i mindre utstrekning gi opplysninger om de enkelte aldersår (generasjoner). Som oftest må vi innskrenke oss til å gi tall for aldersklasser som spenner over flere generasjoner. Dermed blir også valget av klasseintervaller viktig fra et samordningssynspunkt.

En gjennomgang av de 30 personstatistikkene viser at spesifikasjonene etter alder varierer en del fra statistikk til statistikk. I 10 statistikker var det gitt noen opplysninger om hvert enkelt aldersår. 9 statistikker hadde spesifikasjoner på femårige aldersgrupper (0-4, 5-9 o.s.v.), mens 16 statistikker opererte med aldersgrupperinger fremkommet gjennom aggregering av femårsgruppene. Andre aldersklasseinndelinger forekom i 13 av statistikkene. Som en grov sammenfatning kan vi si at aldersopplysningene i om lag to tredjeparter av statistikkene gis for femårsgruppene eller for aggregater av disse. I den resterende tredjepart av statistikkene er hovedutgangspunktet et annet. Dette gjelder i første rekke undervisnings- og kriminalstatistikkene, hvor bestemmelser om skolepliktig alder, om kriminell lavalder o.l. er naturlige utgangspunkt for aldersklasseinndelingen.

Av hensyn til samordningen ville en felles aldersklasseinndeling være ønskelig. Det viser seg imidlertid vanskelig å innpasse de kryssende interesser i en felles aldersklassenorm. Å innføre en bindende aldersklassestandard er derfor neppe formålstjenlig. Derimot må det stilles krav til statistikkene om at disse ordnes på en måte som muliggjør sammenstillinger av hovedtall for visse nærmere fastlagte aldersgrupper. Disse grupper bestemmes i første rekke ut fra behovene for aldersspesifikasjoner i det påtenkte personregnskapsopplegg. Selv om disse behov foreløpig ikke er spesifisert i detalj, foreslås statistikkenes aldersgrupperinger snarest undergitt en kritisk nyvurdering etter følgende retningslinjer:

1) I alle personstatistikker bør det gis visse hovedtall for de enkelte aldersår (generasjoner)

2) Aldersklasseinndelingen skal som hovedregel bygge på femårsgrupper (0-4, 5-9, 10-14, 15-19 o.s.v.) eller på tiårsgrupper (0-9, 10-19, 20-29 o.s.v.)

3) I statistikker hvor det er uhensiktsmessig å følge hovedregelen ovenfor, må opplysninger etter punkt 1 gis i et omfang som muliggjør oppstilling av visse hovedtall for tiårsgrupper

4) Aldersspesifikasjonen i personstatistikkene må dessuten muliggjøre oppstillinger av hovedtall for aldersgruppene:

0-6 år, førskolealder
7-15 år, skolepliktig alder
16-69 (66) år, yrkesaktiv alder
70 (67) år og over, pensjonsalder

Den foreslåtte gjennomgang av statistikkenes aldersklasseinndelinger er ikke ment å skulle medføre radikale endringer i allerede innarbeidd praksis. Samordningen av aldersklasseinndelingene er dessuten en sak som i første rekke er aktuell for opplysninger om alder i fylte år ved årets slutt.

Vedlegg

Retningslinjer for definisjon og klasseinndeling av kjennemerket alder

Direktøren har i planleggermøte 10. januar 1970 fastsatt disse retningslinjer for - definisjon og klasseinndeling - av kjennemerket alder i offentlig norsk statistikk:

§ 1

Kjennemerket alder skal i personstatistikk defineres som alder i fylte år pr. kalenderårets slutt.

§ 2

Oppgaver for aldersklasser skal som hovedregel gis for femårsgruppene 0 - 4, 5 - 9, 10 - 14, 15 - 19 osv. eller for aggregater av disse grupper.

§ 3

Tabellverkene bør, når materialets størrelse tillater det, gi visse grove hovedtall for de enkelte aldersår (generasjoner).

§ 4

Når særlige grunner tilsier det, kan alderen bestemmes som fylte år, måneder, uker eller dager pr. tidspunktet for visse hendinger. Dette gjelder spesielt i tilfeller der en vil belyse i detalj hvorledes hyppigheten av visse begivenheter avhenger av den biologiske alder på hendingstidspunktet.

IM/TK, 14/3-72

Foredrag under BEDRIFTSØKONOMISK UKE 1971,
Geiranger 16.-18. september 1971

DATAGRUNNLAGET FOR UTDANNINGSPOLITIKKEN

Av Idar Møglestue

Innledning

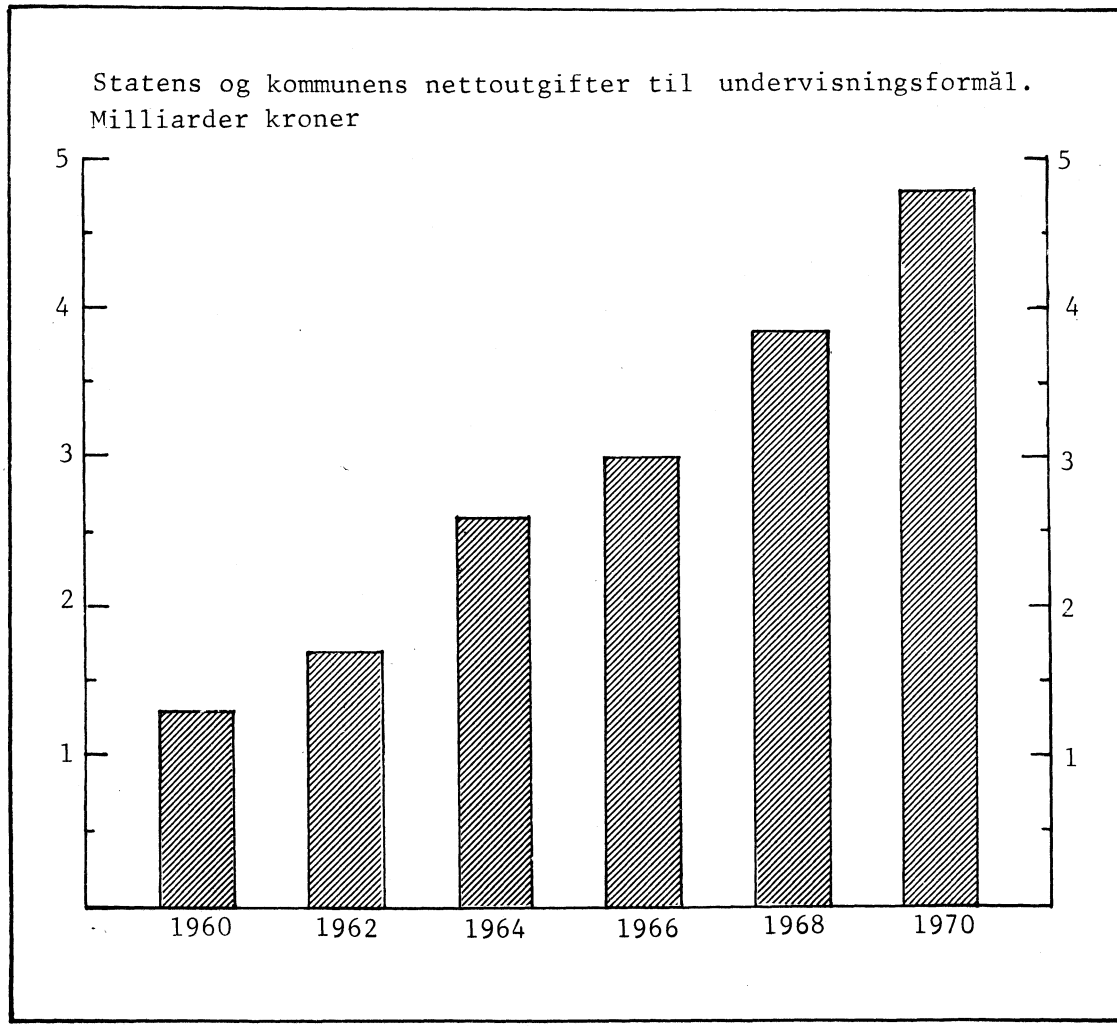
Å sørge for et hensiktsmessig utdanningssystem er blant de viktigste oppgaver våre myndigheter har. De mange planleggingsoppgaver, prioriteringsspørsmål og samfunnsmessige kontrolloppgaver innenfor dette store og stadig voksende offentlige virksomhetsområde, stiller store krav til det dataunderlag som de sentrale og lokale myndigheter må ha for å kunne treffe rasjonelle beslutninger. Det er dette dataunderlag statistikken skal gi.

Jeg kommer her bare til å snakke om den offentlige utdanningsstatistikken. Det betyr imidlertid ikke at andre statistikkområder, data innsamlet som ledd i administrative rutiner og data fra spesialundersøkelser er av underordnet betydning for utdanningspolitikken. Utdanningsstatistikken bør imidlertid oppfattes som grunnstammen i det datamessige informasjonssystem som en trenger for å kunne forme ut en rasjonell utdanningspolitikk.

Tradisjonell utdanningsstatistikk

Vår nåværende utdanningsstatistikk bygger i hovedsak på oppgaver innhentet fra skolene over tallet på elever og lærere pr. 1. oktober hvert år. På oppgavene, som gis på ett skjema for hver skole, er elevtallet spesifisert etter kjønn, alder og klasse. I tillegg er det på skjemaene for enkelte skoleslag bl.a. spurt etter elevenes fordeling etter kurstype og målform. Læreroppgavene gis spesifisert etter kjønn og utdanningsbakgrunn. Dette hovedmønster har et par unntak. For studenter ved universiteter og vitenskapelige høgskoler innhentes det således mer detaljerte opplysninger på ett skjema for hver student. Dessuten er det siden 1968 utarbeidet en statistikk over gymnaseksamen, som bygger på individualoppgaver om artianerne.

Den statistikk som utarbeides etter disse oppgaver, kan vi best få et innblikk i ved å ta for oss noen av hovedtallene i Statistisk Sentralbyrås publikasjoner. La oss først se hva statistikken over stats- og kommune-regnskapene sier om utdanningens finansielle side.



De offentlige utgifter til undervisningsformål økte fra 1,3 milliarder kroner i 1960 til 4,8 milliarder kroner i 1970. Dvs. at de offentlige utdanningsutgifter ble mellom tre- og firedobblet i løpet av siste tiårsperiode. Regnet i fast kroneverdi økte utgiftene nesten 2 1/2 ganger. Av de samlede offentlige budsjetter gikk snaut 15 prosent til utdanningsformål i 1960 mot vel 18 prosent i 1970. Utdannings-utgiftene la i 1970 beslag på vel 28 prosent av kommunebudsjettene og på 14 prosent av statsbudsjettet.

Denne vekst i de offentlige utdanningsutgifter skyldes blant annet en oppgang i elevtallene. Samlet elevtall gikk opp med 113 000 eller med 18 prosent fra 1960 til 1970.

Elevtall pr. 1. oktober

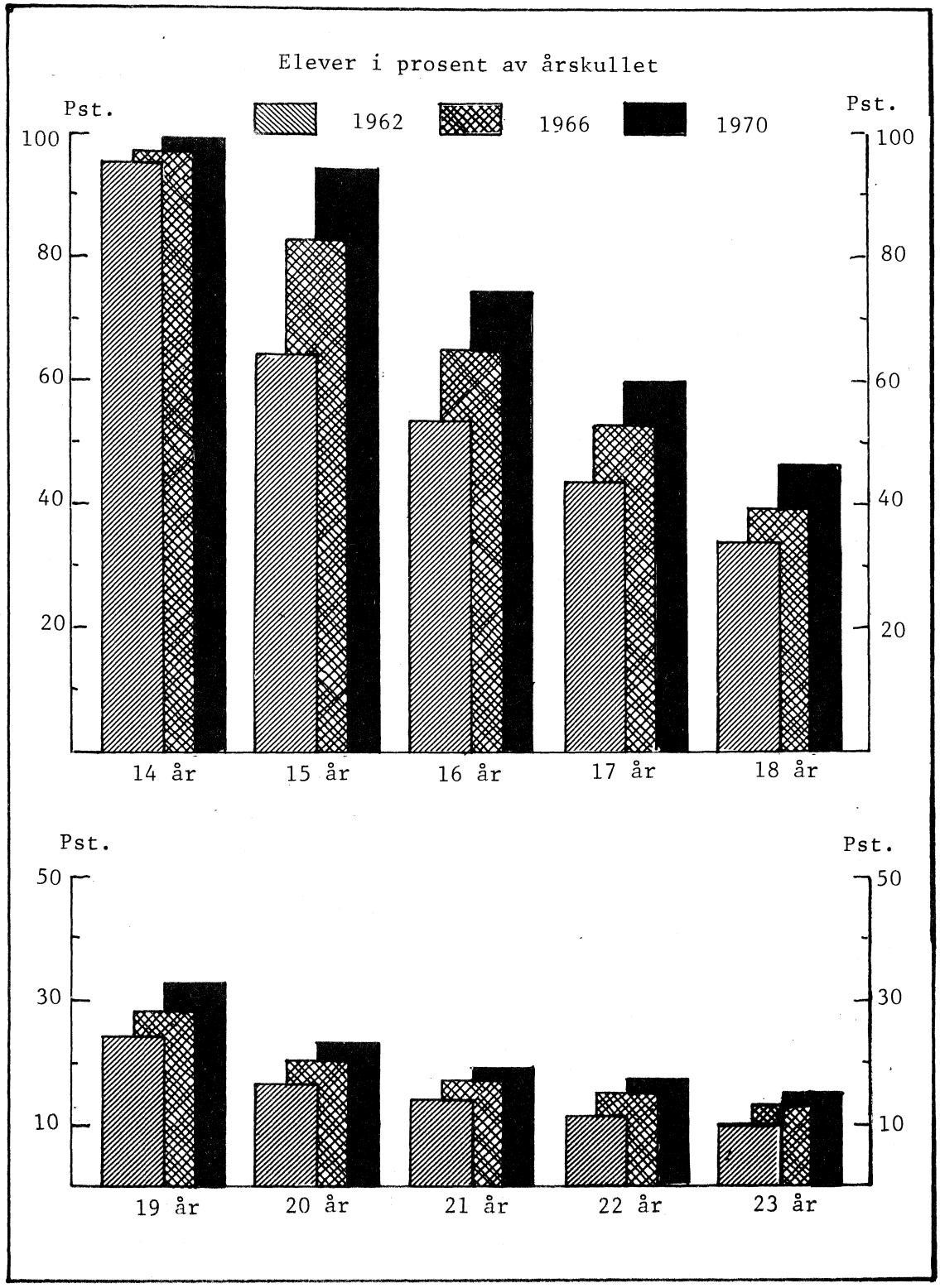
	Grunn-skoler	Videregående allmenn-skoler	Fag- og yrkesskoler	Universitet og høgskoler	I alt
1960	482 242	97 450	50 950	9 254	639 896
1962	479 692	108 545	58 030	12 457	658 724
1964	485 079	112 955	69 569	17 070	684 673
1966	502 091	109 219	75 663	21 001	707 974
1968	525 114	97 280	82 111	241115	728 620
1970	548 794	83 865	89 899	30 461	753 019

Det meste av elevtallsøkingen falt på grunnskolene, som hadde nesten 70 000 flere elever i 1970 enn ti år tidligere. Prosentvis var imidlertid økningen desidert sterkest ved universiteter og høgskoler, hvor studenttallet ble mer enn tredobblet. Også fag- og yrkesskolene hadde en betydelig vekst, i det elevtallet i denne skolegruppen økte med i alt 76 prosent.

For å eliminere den innvirkning ungdomskullenes størrelse har på elevtallene, er det regnet ut elevhyppigheter (elever i prosent av årskullet) for ulike aldersklasser. Disse forholdstall viser, som rimelig er, at utdanningsfrekvensen avtar raskt med stigende alder. Fra 14-års-alderen, hvor så å si samtlige er skoleelever, faller elevhyppigheten til 75 prosent for 16-åringer. Den utgjorde i 1970 33 prosent for 19-åringer og vel 11 prosent for 24-åringer.

I løpet av 1960-årene er elevhyppighetene stadig økt for alle aldersgrupper. Oppgangen var særlig markert for 15-åringene, noe som skyldes utbyggingen av den 9-årige grunnskole.

Jeg nevner også at statistikken forteller om en viss utvikling i retning av større likestilling mellom kjønnene i utdanningssammenheng. Ved universiteter og høgskoler, hvor mannsdomminansen er betydelig, økte således andelen av kvinnelige studenter fra 20 prosent i 1960 til ca. 28 prosent i 1970.


Elever i prosent av årskullet
Pst.
1962
1966
1970
Pst.
100
80
60
40
20
14 år
15 år
16 år
17 år
18 år
Pst.
50
30
10
19 år
20 år
21 år
22 år
23 år

Tallet på lærere tilsatt i full post i skoleverket gikk opp fra 28 665 i 1960 til 43 500 i 1970, dvs. en oppgang i lærertallet på hele 52 prosent. Storparten av lærerøkningen falt på grunnskolene, hvor lærertallet gikk opp med 10 000 i løpet av tiårsperioden. Prosentvis økte imidlertid tallet på lærer- eller forskerstillinger mest ved universitetene og høgskolene og ved fag- og yrkesskolene.

Tallet på lærere har i løpet av siste tiårsperiode økt forholdsvis langt sterkere enn tallet på elever. Antall elever pr. heltidsansatt lærer gikk således ned fra 22,3 i 1960 til 17,3 i 1970. Denne utvikling i forholdstallet mellom elever og lærere gjelder alle skoleslag, bortsett fra universitetene og høgskolene. Ved våre vitenskapelige læreanstalter har tallet på undervisnings-/forskerstillinger ikke holdt tritt med den eksplosive vekst i studenttallet. Ved universiteter og høgskoler er således tallet på studenter pr. lærer gått opp fra 7,6 i 1960 til 12,1 i 1970.

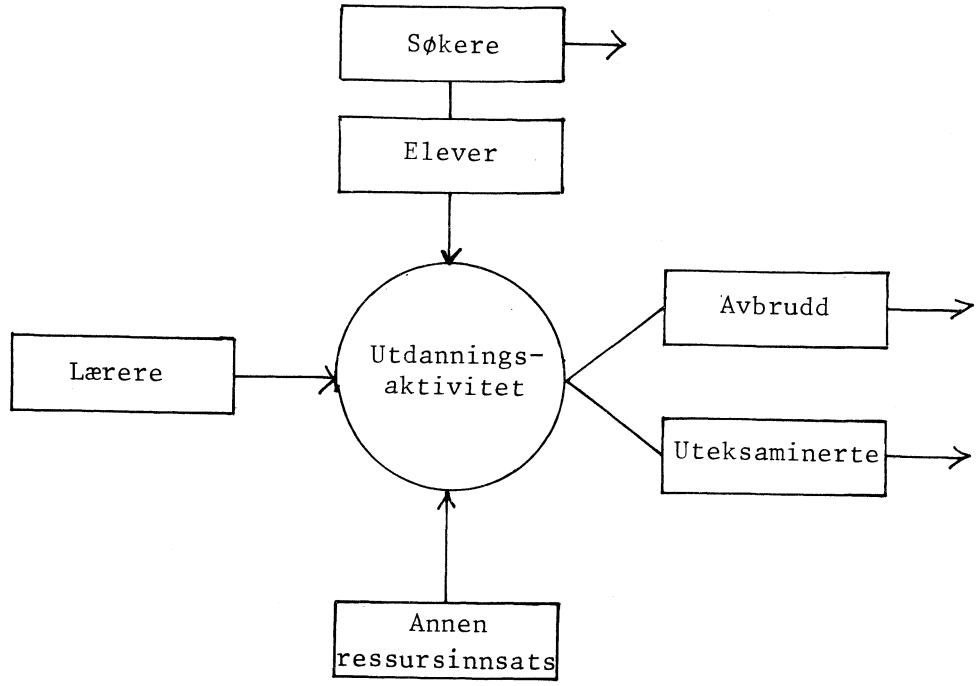
Behov og behovsdekning

Skolemyndighetene har i sitt arbeid med budsjettsaker, undervisningsopplegg, disponering av skolebygg, opplegg av eksamener m.m. behov for løpende oversikter over tallet på søkere, elever, klasser, uteksaminerte og lærere. De trenger også oversikter over kostnadene ved ulike former for undervisning. På bakgrunn av at like muligheter til utdanning uansett bosted, yrke og inntekt er en sentral målsetting i utdanningspolitikken, er det stilt krav om at statistikken må gi opplysninger om elevgruppers bosted og miljømessige bakgrunn ellers. Det er nødvendig at statistikken gir slike opplysninger for blant annet å kunne vurdere om og i hvilken grad utviklingen skjer i ønsket retning. Et annet høyprioritert krav er at utdanningsstatistikken må klarlegge elevgruppers gang gjennom utdanningssystemet. Slike oversikter er en forutsetning for utarbeiding av prognoser over den framtidige etterspørsel etter utdanning. Prognoser av dette slag, som er et viktig grunnlag for utdanningsplanleggingen, vil i dag være svært usikre fordi det ikke foreligger opplysninger om i hvilken grad elevene fullfører en påbegynt utdanning, fortsetter med videre skolegang eller går ut i arbeidslivet etter endt utdanning.

Utdanningsvirksomheten kan oppfattes som sammensatt av en rekke prosesser eller aktiviteter, hvor innsatsfaktorer (elever, lærere, læremidler o.l.) omformes til produkter uttrykt ved kunnskapstilveksten hos

de elever som har deltatt i aktivitetene. Oppfattet på denne måte kan utdanningssystemets enheter spesifiseres som i skissen nedenfor.

UTDANNINGSSYSTEMETS ENHETER



For alle de spesifiserte enheter er det reist krav om innsamling av statistisk erfaringsmateriale hvor enhetene blir karakterisert ved en rekke kjennemerker. Det er også vesentlig at dette materialet ordnes slik at sambandet mellom de ulike aktiviteter innenfor utdanningsvirksomheten er i orden, og at tilknytning til aktuelle data for sektorer utenfor utdanningssystemet kan etableres.

Sammenholder vi vår tradisjonelle utdanningsstatistikk med det som her er sagt om statistikkbehovet vil vi finne at:

1) Enheten utdanningsaktivitet er for dårlig og for usystematisk spesifisert i statistikken. Oppgavene spesifiseres i dag i hovedsak etter en slags skolegruppering.
2) Elevtilgangen er utilstrekkelig registrert. De søker-oppgaver som gis er begrenset til summariske oppgaver, som p.g.a. dobbeltsøknader o.l. ofte er mer villedende enn veiledende. Elevtall registreres bare pr. 1. oktober hvert år, og da med en heller dårlig kjennemerkespesifikasjon.

3) Oppgaver over utdanningsproduktet mangler på det nærmeste helt.
4) Også lærere og annen resursinnsats gis det mangelfulle opplysninger om.
5) Vårt tradisjonelle statistikkopplegg muliggjør heller ikke noen form for sammenkobling av data for ulike utdanningsaktiviteter eller for sammenkobling av oppgaver fra utdanningssystemet med oppgaver for andre aktiviteter i samfunnet.

Alt i alt må vi konstatere at utdanningen er et av de områder hvor kløften mellom behovet for statistikk og tilgangen på statistikk er særlig stor. Viktige utdanningspolitiske avgjørelser må derfor enten treffes på grunnlag av mangelfulle opplysninger eller må baseres på kostbare enkeltundersøkelser.

Utbygging av statistikken

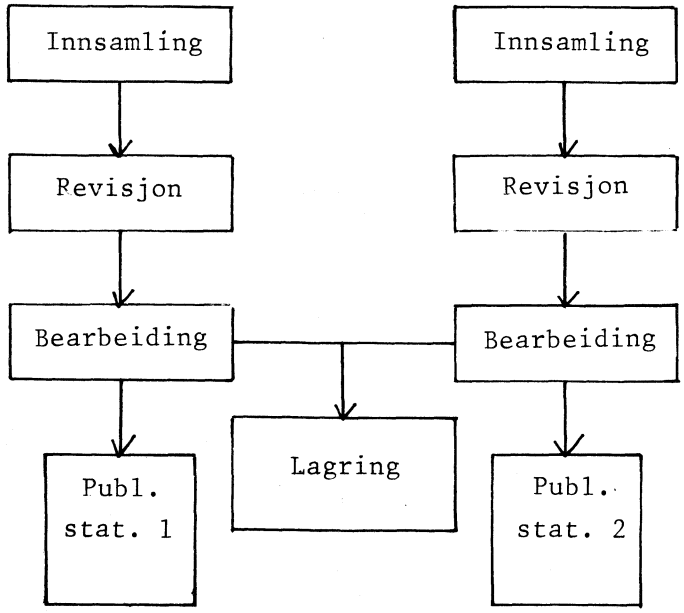
Dette er bakgrunnen for at Statistisk Sentralbyrå, i samarbeid med Kirke- og undervisningsdepartementet og andre statistikkbrukere, nå arbeider med planer for en omfattende utbygging av utdanningsstatistikken.

Utbyggingen vil foregå i etapper. Først tar en sikte på å bygge ut en ny statistikk over all avsluttet utdanning. Gjennomføringen av denne etappen er forlengst påbegynt, da skolene fra og med 1971 sender inn oppgaver til Byrået over personer som avslutter en utdanning. Annen etappe er ment satt ut i livet i 1973, da statistikken over søkere og elever skal bygges ut. Siden kommer turen til statistikken over lærer- og annen resursinnsats.

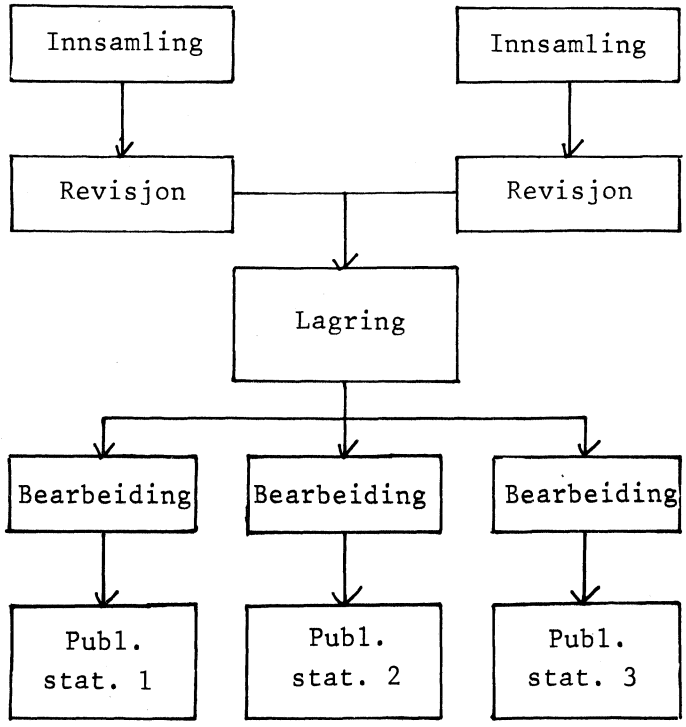
Det som i første rekke skiller det nye statistikkopplegget fra tradisjonelle opplegg, er at statistikken skal utarbeides etter individualoppgaver, dvs. oppgaver hvor registreringsenheten er den enkelte søker, elev eller eksaminand. Det er bare gjennom innhenting av slike oppgaver at statistikkprodusenten kan bli satt i stand til å produsere statistikk som gir en rimelig grad av behovsdekning.

Det andre prinsipielt nye ved de statistikkplaner som foreligger, er at utdanningsstatistikken vil bli utarbeidet etter den såkalte arkivstatistiske metode.

TRADISJONELL METODE



ARKIVSTATISTISK METODE



Det arkivstatistiske system bygger, i motsetning til tellingene, på den grunnidé at opplysninger fra individuelle oppgavegivere arkiveres gjennom lengere tid slik at de er tilgjengelige og kan hentes fram fra dataarkivet ved behov. De tekniske mulighet for dette er skapt ved datamaskinene, som har åpnet adgang til å ordne lagre og gjennomsøke store datamasser på kort tid og med små kostnader.

Et vilkår er at det er etablert et permanent og entydig identifikasjonssystem for de statistiske enheter, slik at vi vet hvilke enheter de lagrede opplysninger refererer seg til. Fødselsnummeret er et slikt identifikasjonssystem for enheter som refererer seg til enkeltpersoner. Innføringen av dette nummeret i elevregistreringen, som nå pågår, er således et nødvendig ledd i arbeidet for å bedre datagrunnlaget for utdanningspolitikken. Det samme gjelder arbeidet med opplegget av en standard for utdanningsgruppering og opprettingen av et register over utdanningsinstitusjoner. De nummer som enkeltutdanningene og skolene er tildelt her, muliggjør en sikker identifikasjon av enheten utdanningsaktivitet.

På dette grunnlag vil de løpende oppgaver som må hentes inn fra skolene, være relativt enkle fortegnelser over søkere, elever og eksaminander. I stor utstrekning tar en sikte på å utnytte oppgaver som alt sendes inn for administrative formål. Gjennom en sammenkobling over tiden av de løpende meldinger vil en likevel få et relativt detaljert erfaringsmateriale for studier av utdanningsaktivitetene og persongruppers utdanningsatferd. Utnytting av de muligheter som foreligger for en supplering av dette erfaringsmateriale med opplysninger fra andre statistiske kilder, vil kunne mangedobble materialets informasjonsverdi. Det er gjennom en slik intergrert innsamling og behandling av data, at utdanningsstatistikken nå vil bli satt i stand til bedre å fylle sin funksjon i det informasjonssystem som trengs for utformingen av en rasjonell utdanningspolitikk.

I et statistikkopplegg basert på innsamling, arkivering og bruk av individualoppgaver, kan en ikke se bort fra at oppgavene kan tenkes nyttet på en måte som er til skade for enkeltpersoner. For å forhindre en slik mulighet, er det utarbeidet spesielle retningslinjer for behandlingen av oppgaver til utdanningsstatistikken. Disse retningslinjer supplerer de alminnelige regler Statistisk Sentralbyrå har for sin egen behandling av konfidensielle opplysninger - regler som har bestått sin prøve ved at Byrået i snart 100 år har behandlet oppgaver av strengt fortrolig art for navngitte personer uten at det har oppstått noen problemer i så måte.

IT/SH/WTD, 9/8-71

Kontroll av data samlet inn ved arbeidskraftundersøkelsene

av

Sverre Hovind og Ib Thomsen

Innhold

KONTROLL AV DATA SAMLET INN VED ARBEIDSKRAFTUNDERSØKELSENE

1. Innledning

Dette notat har to formål:

1. Det skal være en orientering til arbeidsutvalget om hvorledes vi tenker oss kontrollene lagt opp. (For dette formål trenger en bare å lese avsnittene 1-3.)
2. Det skal være et utgangspunkt for sammen med Systemkontoret å lage et system av kontroll- og opprettingsrutiner.

Vi har i notatet foreslått noen automatiske opprettinger i håp om at dette skal hjelpe oss til å redusere produksjonstiden. Av flere grunner har vi i denne omgang bare tatt sikte på automatisk oppretting av spørsmål, som inngår i A-tabeller. Når det gjelder kontroll av spørsmål for B-tabeller foreslår vi at disse utføres på tradisjonell vis. I avsnitt 4 er gitt en oversikt over alle kontrollrutinene. I avsnitt 5 er et utkast til maskinelle kontroller og opprettinger beskrevet, og avsnitt 6 er et utkast til manuelle kontroller. Når kontrollprogrammene er laget bør det foreligge en nøyaktig beskrivelse av disse.

2. Vurdering av den automatiske rutinen.

Av hensyn til framtidige undersøkelser er det nødvendig at de automatiske rutiner vurderes og justeres etter de erfaringer vi gjør ved de to første tellinger. Vi kan tenke oss at vi blir nødt til å sammenlikne tidligere svar med de svar vi har fått ved en undersøkelse. Slike kontroller er ikke tatt med i dette forslag.

Selv om vi regner med at de fleste feil vil bli oppdaget og opprettet automatisk har vi vært nødt til å "kaste ut" noen tilfeller til manuell behandling, se avsnitt 5. Disse skjemaer må da punches opp og legges tilbake i massen etter oppretting. Hvor stor masse det her blir tale om tør vi ikke si noe om på forhånd.

For å vurdere kvaliteten av rutinen må vi telle opp to typer av feil, som kan gjøres:

Feiltype I: De skjemaer, som kastes ut til manuell behandling, men som det er mulig å rette opp ved en enkel justering av rutinen.

Feiltype II: Skjemaer som rettes opp feil.

Dersom antall feil av type I blir stort vil det manuelle arbeid med oppretting ikke kunne gjennomføres innen den fastsatte tid. (Ca. 14 dager.) Hvis antall feil av type II blir stort vil kvaliteten av de publiserte tall bli dårlig.

3. Uoppgitt - Frafall.

Etter maskinell og manuell oppretting må vi regne med ugyldige svar. Dette kan f.eks. skyldes feil utfylt skjema e.l. Det kan skyldes at tiden ikke tillater å fullføre den manuelle oppretting. Vi foreslår da følgende:

1. Alle skjemaer med manglende alder og/eller kjønn tas ut og regnes som frafall. Likeledes alle skjemaer, som ikke kan plasseres i en av kategoriene i tabell 1.
2. Skjemaer som bare mangler en enkelt av de øvrige opplysninger kan gå inn i de berørte tabeller under kategorien uoppgitt. Årsaken til dette er å forenkle oppblåsingen. Se notat om oppblåsing.

4. Oversikt over hele kontrollrutinen.

Innsamlings- og bearbeidingsprosessen kan inndeles i følgende rutiner:

1. Intervjuingen
2. Mottakelsen av skjemaer
3. Dataoverføring
4. Maskinell og manuell kontroll og oppretting
5. Produksjon av A-tabeller

Etter at A-tabellene er publisert gjennomgås bearbeidingsprosessen forfra med hele skjema.

(i) Manuell revisjon av skjema

(ii) Opprettede skjema (altså bare de med feil) sendes til overføring og samsorteres inn i A-tabell filen

(iii) Maskinell kontroll og oppretting.

- a. Validitetskontroller
- b. Kombinasjonskontroller
- c. Grensekontroller.

(iv) Oppretting

(v) Produksjon av B-tabeller.

I tillegg til dette kommer vurdering av kontroll og opprettingsrutinen for A-tabellene.

5. Beskrivelse av maskinell kontroll og opprettingsrutine ved produksjon av A-tabeller.

Prinsippet for opprettingen er at hvis en kode er ugyldig i et felt leses skjema for å finne hvilken gyldig kode, som er mest sannsynlig. For noen av feltene testes også om ét gyldig svar er konsistent med andre svar på skjema. Hvis vi ikke kan finne hvilken kode som er mest sannsynlig eller hvis skjema er grovt inkonsistent tas det ut for manuell oppretting.

5.1 Kontroll av yrke, næring og yrkesstatus.

Denne del av kontrollen er karakterisert ved at ingen maskinell oppretting foretas. Det vil kanskje vise seg at antall feil av type I vil bli meget stort og at det er mulig å rette noen opp automatisk, men vi har valgt å satse utelukkende på manuell oppretting. Vi foreslår både validitetskontroller og kombinasjonskontroller. I tabell 1 er gitt alle gyldige kombinasjoner av yrke, næring og yrkesstatus.

5.2 Kontroll av alder, kjønn og ekteskapelig status.

Også ved denne kontroll foreslås manuell oppretting. Alle records med blank i fødselsår eller kjønn tas ut for manuell oppretting.
Vi regner med at disse to opplysninger vil bli kontrollert ved mottakelsen av skjema slik at denne massen ikke blir for stor.

a. Kontroll av alder: Kontroll for riktig aldersgruppe (16-74). Hvis feil: Ta ut skjema.

b. Kontroll av kjønn: Hvis ugyldig kode tas skjema ut.

c. Kontroll av ekteskapelig status: Hvis ugyldig kode tas skjema ut.

d. Kombinasjonskontroller: Hvis gift/før gift, kvinne og under 18 år tas kortet ut.

Tabell 1

Næringsområde	Yrke	Yrkesstatus
11-96	00	-
11-96	01	-
11-96	02	-
93	03	-
93	04	-
35, 62, 91, 93, 95	05	-
93	06	-
91,93	07	2
83,91	08	-
11-96	09	-
11-96	0x	-
91	10	2
11-96	11	-
11-96	20	-
11-96	21	-
11-96	29	-

Nærings-område	Yrke	Yrkes-status
61,62	30	-
34, 61-83	31	-
21-62, 81-83	32	-
21-39, 61,62	33	-
11, 12, 93	40	-
11, 20, 93	41	-
12	42	-
13	43	-
12, 33, 34	44	-
21, 23, 29, 50	50	-
21, 23, 29, 50	51	-
21, 23, 29	52	-
21, 23, 29	59	-
71	60	-
71	61	-
71	62	-
71	63	2
11-96	64	-
71	65	2
71	66	2
11-96	67	2
11-96	68	2
71	69	2
32	70	-
32, 33, 38, 62	71	-
32,35	72	-
37	73	-
38, 39, 61, 62, 93	74	-
21-71, 95	75	-
11-96	76	-
11-96	77	-
11-96	78	-
29,50	79	-
32, 34, 38	80	-
36	81	-
31, 61, 62, 63	82	-
31, 34, 35, 36, 37, 38, 41	83	-
31	84	-
11-96	85	-
11-96	86	-
11-96	87	-
11-96	88	-
11-96	89	-
11-96	90	-
11-96	91	-
63,71	92	-
11-96	93	-
93,95	94	-
11-96	95	-
94	96	-
34, 83, 95	97	-
95	98	-
11-96	99	-
91	x1	2
11-96	x2	-

5.3 Deling av massen i sysselsatte og resten.

For å plassere en person som sysselsatt bruker vi spørsmål 1, spørsmål 2 samt opplysninger om yrke og næring.

I spørsmål 1 har vi tre muligheter:

Kode 1
" 2
Ugyldig kode

I spørsmål 2 deler vi inn i:

Større enn null
Null

Yrke og næring inndeles i:

Oppgitt
Uoppgitt

Dette gir 24 mulige kombinasjoner. Vi skal i det følgende plassere hver person i kategoriene:

1. Sysselsatt
2. Personer uten inntektsgivende arbeid
3. Personer som må tas ut for manuell oppretting.

1 i kol. 26 og større enn null i (27-28) kat. 1

1 i kol. 26 og null i (27-28) gjøres følgende:

Hvis oppgitt i næring eller yrke kat. 1

Hvis uoppgitt i næring og yrke kat. 3

2 i kol. 26 og større enn null i (27-28) og gyldig i yrke eller næring kat. 1

2 i kol. 26 og større enn null i (27-28) og ugyldig i yrke og næring kat. 3

2 i kol. 26 og null i (27-28) gjøres følgende:

Hvis gyldig kode i kol. 39 kat. 2

Hvis ugyldig kode i kol. 39 kat. 3

Ugyldig i kol. 26 og større enn null i (27-28) og gyldig i yrke eller næring kat. 1

Ugyldig i kol. 26 og større enn null i (27-28) og ugyldig yrke og næring kat. 3

Ugyldig i kol. 26 og null i (27-28) gjøres følgende:

Hvis gyldig kode i kol. 39 kat. 2

Hvis ugyldig kode i kol. 39 kat. 3

5.4 Oppdeling av personer uten inntektsgivende arbeid.

Massen under kategori 2 på foregående side skal nå deles i følgende delmasser:

1. Arbeidsaktive. Midlertidig fraværende
2. " . Personer som søker arbeid
3. " . Studenter, skoleelever
4. " . Vernepliktige
5. " . Personer i husarbeid hjemme
6. " . Andre
7. Ikke arbeidsaktive. Personer som søker arbeid
8. " . Andre
9. Personer som tas ut for manuell oppretting.

Massen deles på grunnlag av innholdet i kol. 39, 40, 41, 42 og 43.

0 eller 9 i kol. 39 og gyldig i kol. 40 kat. 1

0 eller 9 i kol. 39 og ugyldig i kol. 40 kat. 9

1 i kol. 39 og gyldig yrke eller næring kat. 1

1 i kol. 39 og ugyldig yrke og næring kat. 9

2 i kol. 39 og 1 i kol. 42 kat. 2

2 i kol. 39 og 2 i kol 42 kat. 5

2 i kol. 39 og ugyldig i kol. 42 gjør følgende:

Hvis gyldig i kol. 43 kat. 2

Hvis ugyldig i kol. 43 kat. 5

3 i kol. 39

Her tas samme kontroll som for husmødre

4 i kol. 39 og gyldig i kol. 41 eller 42 eller 44 kat. 9

4 i kol. 39 og ugyldig i kol 41, 42 og 44 kat. 4

5 i kol. 39

Her tas samme kontroll som for vernepliktige ovenfor.

6 i kol. 39

Her tas samme kontroll som for husmødre

7 i kol. 39 og 1 i kol. 44 kat. 7

7 i kol. 39 og 2 i kol. 44 kat. 8

7 i kol. 39 og annet i kol. 44 gjøres følgende:

Hvis gyldig i kol. 45 kat. 8

Hvis ugyldig i kol. 45 men større enn null i kol. 50-51 kat. 7

Hvis ugyldig i kol. 45 og 50-51 kat. 9

8 i kol. 39 og 1 i kol. 42 kat. 7

8 i kol. 39 og 2 i kol. 42 kat. 5

8 i kol. 39 og ugyldig i kol. 42 gjøres følgende:

Hvis gyldig i kol. 43 kat. 7

Hvis ugyldig i 43 men gyldig i kol. 45 kat. 5

Hvis ugyldig i 43 og ugyldig i kol. 45 kat. 9

6. Utkast til manuell revisjon av skjema etter produksjon av A-tabellene

På lengre sikt bør vi vel kunne komme fram til et kontrollopplegg slik at denne revisjon kan unngås. Vi mener dog at vi i første omgang bør gå skjemaene nøye igjennom både for å veilede intervjuerne og for å vurdere og justere kontrollprogrammene. Revisjonsinstruksen er av tradisjonell art. Nedenfor er gitt et utkast til en henvisningskontroll.

Henvisningskontroll

Kolonne	Spørsmål	Svaralternativ
3- 5		Utvalgsområde skal være 3-sifret 001-286
6- 9		År, kvartal og antall ganger IO har vært med - 4-sifret
10-13		Husholdningsnr. skal være 4-sifret
14		Husholdningsmedlemsnr. skal være 1-sifret
15-25		Fødselsnr. Personnr. skaffes fra personregisteret dersom det ikke alt er påført
26	1	X i boks 1 eller 2
27-28	2	Hvis X i boks 1 kol. 26, skal tallet på arbeidstimer være oppgitt
	3	Dette er kontrollspørsmål om oppgitt timetall er korrekt. Svar under spm. 3 punches ikke
29	4	Hvis X i boks 1 kol. 26, skal det være X i én (og bare én) av boksene 1-4
30-31	5	Hvis X i boks 1 kol. 29, skal det timetallet som IO i alt kunne tenke seg å arbeide, være oppgitt her
32	6	Hvis X i boks 1 kol. 29, skal det være X i én (og bare én) av boksene 1-5
33	7	Hvis X i boks 1 eller 3 kol. 29, skal det være X i én (og bare én) av boksene 1-8
34	8	Hvis X i boks 1 eller 3 kol. 29, skal det være X i boks 1 eller 2
35-36	9	Hvis X i boks 4 kol. 29, skal det timetallet som IO i alt kunne tenke seg å arbeide, være oppgitt her
37	10	Hvis X i boks 4 kol. 29, skal det være X i én (og bare én) av boksene 1-6
38	11	Hvis X i boks 4 kol. 29, skal det være X i én (og bare én) av boksene 1-3

Kolonne	Spørsmål	Svaralternativ
39	12	Hvis X i boks 1 kol. 26, skal det være X i boks 1. Hvis X i boks 2 kol. 26,skal det være X i én (og bare én) av boksene 2-8
40	13	Hvis X i boks 1 kol. 39, skal det være X i én (og bare én) av boksene 0-9
41	14	Hvis X i boks 0, 2, 3, 4, 6, 7, 8 eller 9 i kol. 40, skal det være X i boks 1 eller 2
42	15	Hvis X i boks 2, 3, 6 eller 8 i kol 39, skal det være X i boks 1 eller 2
43	16	Hvis X i boks 1 kol. 42, skal det være X i én (og bare én) av boksene 1-5
44	17	Hvis X i boks 7 kol. 39, skal det være X i boks 1 eller 2
45	18	Hvis X i boks 2 kol. 42 eller kol. 44, skal det være X i én (og bare én) av boksene 1-8
46	19	Hvis X i boks 1, 3, 4 eller 8 kol. 45, skal det være X i boks 1 eller 2
47	20	Hvis X i boks 1 kol. 46, skal det være X i én (og bare én) av boksene 1-6
48-49	21	Hvis X i én av boksene kol. 47, skal det timetallet som IO kunne tenke seg å arbeide pr. uke,være oppgitt her
50-51	22	Hvis X i boks 1 kol. 41 eller kol. 44, eller hvis X i én av boksene 1-5 kol. 43, skal det timetallet som IO kunne tenke seg å arbeide pr. uke, være oppgitt her
52	23	Hvis X i boks 1 kol. 34, eller at det er oppgitt timetall arbeidstid i kol 50-51, skal det være X i én (og bare én) av boksene 1-4
53-54	24	Hvis X i én av boksene 1-4 kol. 52, skal tallet på uker, som IO har søkt arbeid, være oppgitt her
	25	Bedriftens navn og adresse har en behov for å kjenne, når næringsgruppering skjer ved hjelp av bedrifts- og foretaksregisteret
51	25	For personer som søker arbeid for første gang skal det settes ring rundt kode 1
56-57	26	Her kodes næringsområde (to-siffernivå). Bruk vedlagte kodeliste "Standard for Næringsgruppering" Arbeidskraftundersøkelser
58	27	Det skal være X i boks 1 eller 2
59	28	Det skal være X i boks 1 eller 2
60	29	Det skal være X i boks 1, 2 eller 3
61-62	30	Her kodes yrkesområde (to-siffernivå). Bruk vedlagte kodeliste "Alfabetisk register for yrkesklassifisering"
63-64	31	Her skal bare hele år være notert. For mindre enn 1 år skal det stå 00. Hvis tallet på år er færre en 10 settes 0 i første rubrikk f.eks. (06)
65	32	Her skal det være X i en av boksene 1, 2 eller 3. Hvis uoppgitt, vurdér mot andre opplysninger, yrke, alder osv.

Kolonne	Spørsmål	Svaralternativ
66-67	33	Her skal tallet på personer som bor i leiligheten være oppgitt. Hvis tallet på personer er under 10 settes 0 i første rubrikk f.eks. (06)
68-69	34	Hvis tallet på personer i kol. 66-67 er 1, skal det være X i boks 1 eller 2. Hvis det er X i boks 1, skal tallet på syke være oppgitt
70	35	Her skal tallet på barn under 16 år stå oppført
71-72	36	Her skal alder på yngste barn stå oppført
73	37	Her skal det være X i boks 1 eller 2
74	38	Her skal det være X i boks 1 eller 2
75-76	39	Her skal timetallet være oppgitt i hele timer

Vedlegg 1

STATISTISK SENTRALBYRÅ
Kontoret for intervjuundersøkelser
Oslo-Dep., Oslo 1
Tlf. 41 38 20, 41 36 60

ARBEIDSKRAFTUNDERSØKELSE

Prosjekt nr. 3 0 (1- 2)
Utv.omr.nr. (3- 5)
År-Kvartal-Gang (6- 9)
Husholdningsnr. (10-13)
Hush.medl.nr. (14)

Navn _____

Adresse _____

Fødselsdag-mnd.-år For Byrået (15-25)

Spørsmål	Svar	→ = Gå til spm.nr.
1. Utførte De inntektsgivende arbeid av minst en times varighet i forrige uke? (26)	1 Ja 2 Nei	→ 2 → 12
2. Hvor mange timer arbeidde De i forrige uke? (27-28)	Antall timer	
3. Har De da tatt med eventuell overtid og ekstraarbeid, også ekstraarbeid hjemme i forbindelse med arbeidet?	Ja Nei	→ 4 → 2
4. Hvis De i forrige uke kunne ha valgt arbeidstidens lengde, men med samme lønn eller fortjeneste pr. arbeidstime, ville De da ha valgt lengre arbeidstid, samme arbeidstid eller kortere arbeidstid enn den De hadde i forrige uke? (29)	1 Lengre arbeidstid 2 Samme arbeidstid(30 t eller over på spm.2) 3 Samme arbeidstid (under 30 t pr.uke) 4 Kortere arbeidstid	→ 5 → 25 → 7 → 9
5. Hvor mange timer kunne De tenkt Dem å arbeide i alt i forrige uke? (30-31)	Antall timer	
6. Hva var den viktigste grunnen til at De kunne ha tenkt Dem å arbeide flere timer i forrige uke? (32)	1 Har for liten inntekt 2 Har tid til overs 3 Meget interessert i arbeidet 4 For å få mer ansvarsfullt arbeid 5 Annen grunn (spesifiser):	
7. Hva var den viktigste grunnen til at De ikke arbeidde flere timer i forrige uke? (33)	1 Driftsinnskrenkninger 2 Arbeidsstans på grunn av tekniske forhold eller værforhold 3 Arbeidskonflikt 4 Kunne ikke få mer arbeid av andre grunner enn alt. 1-3 5 Husarbeid hjemme 6 Egen sykdom eller skade 7 Sykdom i hjemmet 8 Annen grunn (spesifiser):	
8. Forsøkte De å få mer inntektsgivende arbeid i forrige uke? (34)	1 Ja 2 Nei	→ 23 → 25
9. Hvor mange timer kunne De tenkt Dem å arbeide i alt i forrige uke? (35-36)	Antall timer	
10. Hva var den viktigste grunnen til at De kunne ha tenkt Dem å arbeide færre timer i forrige uke? (37)	1 Svekket helse 2 Skattemessig årsak 3 Ønsket å ha mer tid til fritidssysler 4 Husarbeid hjemme 5 Skolegang, studium 6 Annen grunn (spesifiser):	
11. Hva var den viktigste grunnen til at De ikke arbeidde færre timer i forrige uke? (38)	1 Ikke mulig å få kortere arbeidstid 2 Ikke mulig p.g.a. sesong i virksomheten 3 Annen grunn (spesifiser):	→ 25 → 25 → 25
12. Hva gjorde De hovedsakelig i forrige uke? (39)	1 Var midlertidig fraværende fra inntektsgivende arbeid 2 Utførte husarbeid hjemme 3 Gikk på skole, studerte 4 Var inne til 1.gangs militær- eller sivilarbeidstjeneste 5 Var arbeidsufør 6 Var pensjonist - sluttet i arbeid 7 Var uten arbeid 8 Opptatt med annet (spesifiser):	→ 13 → 15 → 15 → 32 → 32 → 15 → 17 → 15
13. Hva var den viktigste grunnen til at De var midlertidig fraværende fra arbeid i forrige uke? (40)	0 Skolegang, studier 1 Permisjon p.g.a. svangerskap 2 Driftsinnskrenkninger 3 Arbeidsstans p.g.a. tekniske forhold eller værforhold 4 Arbeidskonflikt 5 Repetisjonsøvelse 6 Ferie 7 Egen sykdom eller skade 8 Sykdom i hjemmet 9 Annen grunn (spesifiser):	→ 14 → 25 → 14 → 14 → 14 → 25 → 14 → 14 → 14
14. Forsøkte De å få inntektsgivende arbeid i forrige uke? (41)	1 Ja 2 Nei	→ 22 → 25
15. Forsøkte De å få inntektsgivende arbeid i forrige uke? (42)	1 Ja 2 Nei	→ 16 → 18
16. Hva var den viktigste grunnen til at De forsøkte å få inntektsgivende arbeid i forrige uke? (43)	1 Hadde behov for å tjene penger 2 Hadde tid til overs 3 For å få bedre kontakt med andre mennesker 4 For å gjøre bruk av min utdanning 5 Annen grunn (spesifiser):	→ 22 → 22 → 22 → 22 → 22
17. Forsøkte De å få inntektsgivende arbeid i forrige uke? (44)	1 Ja 2 Nei	→ 22 → 18

18. Hva var den viktigste grunnen til at De ikke søkte inntektsgivende arbeid i forrige uke? (45)
1 ☐ Opptatt av husarbeid hjemme —> 19
2 ☐ Gikk på skole, studerte —> 32
3 ☐ Regnet ikke med å få arbeid innen yrket eller i nærheten av bostedet —> 19
4 ☐ Mangler nødvendig utdanning eller yrkesopplæring —> 19
5 ☐ Arbeidsgiverne synes at jeg er for gammel —> 32
6 ☐ Arbeidsgiverne synes at jeg er for ung —> 32
7 ☐ Synes selv jeg er for gammel —> 32
8 ☐ Annen grunn (spesifiser): ---------- —> 19

19. Ville De ha forsøkt å skaffe Dem inntektsgivende arbeid dersom det hadde vært passende arbeid på eller i nærheten av bostedet? (46)
1 ☐ Ja —> 20
2 ☐ Nei —> 32

20. Hva er den viktigste forutsetning for at De skulle kunne ta inntektsgivende arbeid? (47)
1 ☐ Pass av barna
2 ☐ Hjemmehjelp for eldre
3 ☐ Bedre skatteforhold
4 ☐ Yrkesopplæring/voksenopplæring
5 ☐ At deltidsarbeid kunne oppnås
6 ☐ Annen forutsetning (spesifiser):
---

21. Hvor mange timer kunne De tenke Dem å arbeide pr. uke i den nærmeste framtid? (48-49) ☐☐ Antall timer —> 32

22. Hvor mange timer kunne De tenke Dem å arbeide pr. uke i den nærmeste framtid? (50-51) ☐☐ Antall timer

23. På hvilken måte forsøkte De å skaffe Dem arbeid? (52)
1 ☐ Kontaktet arbeids/sjømannskontoret
2 ☐ Svarte på annonse/annonserte selv
3 ☐ Kontaktet arbeidsgiver
4 ☐ Annen måte (spesifiser)
---

24. Hvor mange uker er det siden De begynte å søke arbeid? (53-54) ☐☐ Antall uker

25. Hvor arbeidde De hovedsakelig i forrige uke:
a. For midlertidig fraværende fra arbeidet bes oppgitt det arbeidssted hvor vedkommende har sitt inntektsgivende arbeid.
b. For arbeidssøkere (personer som svarte ja på spm.15 eller 17) bes oppgitt hvor de sist hadde arbeid.
c. For dem som aldri har utført inntektsgivende arbeid.

Bedriftens navn:

Bedriftens adresse:

--- —> 26

(55) 1 —> 32

26. Hva slags virksomhet er det?

Virksomhetens art:

(56-57) ☐☐ Næringskode

27. Er dette privat eller offentlig virksomhet? (58)
1 ☐ Privat —> 28
2 ☐ Offentlig —> 30

28. Er dette et personlig firma eller er det et aksjeselskap e.l. (59)
1 ☐ Personlig firma
2 ☐ Aksjeselskap e.l.

29. Arbeidde De som selvstendig, som ansatt eller som familiemedlem uten fast avtalt lønn? (60)
1 ☐ Selvstendig
2 ☐ Ansatt
3 ☐ Familiemedlem

30. Hva var Deres HOVEDYRKE i denne virksomheten?
Hovedyrke:

(61-62) ☐☐ Yrkeskode

31. Hvor lenge er det siden De begynte å arbeide i denne virksomheten? (63-64) ☐☐ Antall år

32. Er De ugift, gift eller før gift? (65)
1 ☐ Ugift
2 ☐ Gift
3 ☐ Før gift

De etterfølgende spørsmål stilles bare til en person i husholdningen, men svarene overføres til intervjuskjemaene for de resterende medlemmer av husholdningen.

33. Hvor mange personer bor det i leiligheten? (66-67) ☐☐ Antall personer —> 34
Dersom én person —> 38

34. Er noen av disse syke eller uføre? (68-69)
1 ☐ Ja
2 ☐ Nei
☐ Antall

35. Hvor mange barn under 16 år er det i leiligheten? (70) ☐ Antall

36. Hvor gammelt er det yngste barnet? (71-72) ☐☐ Antall år

37. Var barnet/noen av barna i barnehage eller på daghjem i forrige uke? (73)
1 ☐ Ja
2 ☐ Nei

38. Hadde De noe leid hjelp til husarbeidet i forrige uke? (74)
1 ☐ Ja —> 39
2 ☐ Nei —> 40

39. Hvor mange timer hadde De leid hjelp i forrige uke? (75-76) ☐☐ Antall timer

FYLLES UT AV INTERVJUEREN

40. På hvilken måte ble intervjuet foretatt? (77)
1 ☐ Ved personlig kontakt
2 ☐ Pr. telefon

41. Hvor mange ganger kontaktet De IO? (78)
1 ☐ En gang
2 ☐ To ganger
3 ☐ Tre ganger
4 ☐ Mer enn tre ganger

42. Hvem har De fått opplysningene fra til utfylling av skjemaet? (79)
1 ☐ IO selv
2 ☐ IO's ektefelle
3 ☐ IO's sønn/datter
4 ☐ IO's far/mor
5 ☐ Annen person

JD/JMH/WA, 13/10-72

Variansberegninger ved intervjuundersøkelser[x].

VIII. Reduksjon av parametre i variansformelen for to-trinns utvalg. En teoretisk diskusjon av "design-effekten".

av

John Dagsvik

Innhold

x) De øvrige notatene er samlet i et eget Arbeidsnotat.

1. Innledning

Det er ønskelig å finne en enkel tilnærmelse til den eksakte variansformelen, blant annet for å kunne vurdere hvordan variansen avhenger av de parametre som inngår. Det er spesielt viktig å kunne anslå hvor store befolkningsgrupper vi kan gi data for ut fra et gitt krav om usikkerheten for disse data. Den metode som har vært brukt hittil er å anslå variansen til å være 1,5 multiplisert med variansformelen for rent tilfeldig utvalg. Det finnes imidlertid, så vidt man vet, ingen teoretisk begrunnelse for denne metoden, noe som må sies å være utilfredsstillende. Dette notat tar sikte på å finne et forenklet variansuttrykk basert på teoretiske betraktninger. Det har tidligere [1] vært gjort beregninger over hvor store befolkningsgrupper man kan gi data for. Disse resultatene sammenliknes med overslag basert på de formlene som utledes her, og viser hvilke utslag de a priori antagelser medfører. Vi studerer spesielt "design-effekten", som er forholdet

$$\text{var } \hat{\bar{a}} \;/\; \{\frac{N^2}{n}\, p(1-p)\}.$$

Analysen i dette notatet er basert på resultatene fra den "klassiske" teorien og gjelder derfor ikke uten videre for Byråets utvalgsplan. (Se Notat IV.)[1)]

2. Stratum-variansen i det binære tilfelle

Vi vil betrakte situasjonen hvor $a_i(j, k)$ antar verdiene 0 eller 1. Ifølge Notat (I.6) og (II.3) er

$$(2.1) \qquad \text{Var } \hat{a}_i = \frac{1 - b_i}{b} \sum_j N_i(j)\, \sigma_i^2(j) + \frac{M_i - m_i}{b}\, b_i\, \sigma_i^2,$$

hvor $\sigma_i^2(j)$ blir

1) Nødvendige utdrag gjengis som appendiks nedenfor.

$$(2.2) \qquad \sigma_i^2(j) = \bar{a}_i(j)\,\{1 - \bar{a}_i(j)\}$$

når

$$\frac{N_i(j)}{N_i(j)-1} \approx 1.$$

I Byråets utvalgsplan er størrelsen av primærområdene forutsatt å være omtrent like, slik at vi som en tilnærmelse vil anta at

$$N_i(j) = N_i \,/\, M_i = \bar{N}_i.$$

Ved å utnytte dette kan vi skrive (2.1) som

$$(2.3) \qquad \operatorname{Var} \hat{a}_i = \frac{1-b_i}{b}\, a_i + \left[\frac{M_i-m_i}{M_i-1}\,\frac{b_i}{b} - \frac{1-b_i}{b\,\bar{N}_i}\right] \sum_j a_i^2(j) - \frac{M_i-m_i}{M_i(M_i-1)}\,\frac{b_i}{b}\, a_i^2.$$

La ρ_{ij} og K_i være definert ved

$$a_i(j) = \rho_{ij}\, a_i,$$

$$K_i = \sum_{j \geq 1} \rho_{ij}^2.$$

Vi vil kalle K_i den primære homogenitetsparameter for stratum nr. i.

Begrunnelsen for denne betegnelsen er at K_i øker jo større forskjell det er mellom primærområdene i stratumet. Dette vil bli nærmere drøftet i avsnitt 5.

La $\underset{\sim}{\rho}_i$ være linje vektoren

$$\underset{\sim}{\rho}_i = (\rho_{i1}, \rho_{i2}, \ldots, \rho_{iM_i}).$$

Vi ser at når $a_i > 0$, varierer $\underset{\sim}{\rho}_i$ i området

$$A_i = \{\underset{\sim}{\rho}_i \mid \sum_j \rho_{ij} = 1, \quad 0 \leq \rho_{ij} \leq \bar{N}_i \,/\, a_i\}$$

fordi $a_i(j) \leq N_i(j) = \bar{N}_i$. Ved å innføre K_i i likning (2.3), får vi

$$(2.4) \qquad \operatorname{Var} \hat{a}_i = \frac{1-b_i}{b}\, a_i\left(1 - \frac{a_i}{N_i}\right) + a_i^2\left(\frac{M_i-m_i}{m_i(M_i-1)} - \frac{1-b_i}{N_i b}\right)(M_i\, K_i - 1).$$

3. Total-varians

Oslo har ett stratum hvor utvalget trekkes rent lotterisk. I (II.5) er variansen til estimatoren $\hat{a}_0$ funnet å være lik

$$(3.1) \qquad \operatorname{Var} \hat{a}_0 = \frac{1-b}{b} N_0 \sigma_0^2 = \frac{1-b}{b} a_0 (1 - \frac{a_0}{N_0}).$$

Den totale varians er variansen til $\hat{\tilde{a}} = \hat{a}_0 + \hat{a}$, altså

$$\operatorname{Var} \hat{\tilde{a}} = \operatorname{Var} \hat{a}_0 + \sum_{i \geq 1} \operatorname{Var} \hat{a}_i .$$

I utvalgsplanen varierer M_i mellom 30 og 38 unntatt for Bergen og Trondheim (BT), der M_i ligger mellom 15 og 18. Det gjennomsnittlige antall primærområder for hele landet, unntatt (BT), er altså $M = 34$. Dersom vi benytter samme nummerering av områdene som i [2], har (BT) numrene fra 40 til 46. Oslo lar vi fortsatt ha nummer 0. I de videre utledninger vil vi benytte tilnærmelsen

$$M_i = M = 34 \qquad \text{for } 1 \leqq i \leqq 40,$$

$$M_i = \tfrac{1}{2}M = 17 \qquad \text{for } 41 \leqq i \leqq 46.$$

For Byråets utvalgsplan for 300 intervjuere gjelder

$$m_i = m = 6 \qquad \text{for } 1 \leqq i \leqq 40,$$

$$m_i = \tfrac{1}{2}m = 3 \qquad \text{for } 41 \leqq i \leqq 46.$$

Siden vi har antatt at primærområdene er like store, vil tilsvarende gjelde for populasjonen i strataene, dvs.

$$N_i = N_1 \qquad \text{når } 1 \leqq i \leqq 40,$$

$$N_i = \tfrac{1}{2}N_1 \qquad \text{når } 41 \leqq i \leqq 46.$$

La F være slik at følgende ulikheter er oppfylt.

$$(3.2) \qquad \begin{aligned} & MK_i - 1 \leqq F, && \text{for } 1 \leq i \leq 40, \\ & (\frac{M-m}{m(M-2)} - \frac{1-b_1}{N_1 b})(MK_i - 2) - \frac{1-b_1}{bN_1} \leqq (\frac{M-m}{m(M-1)} - \frac{1-b_1}{N_1 b})F && \text{for } 41 \leq i \leq 46. \end{aligned}$$

Siden $b_i = \frac{M}{m} b = b_1$, vil et slikt valg av F medføre at

$$(3.3) \qquad \operatorname{Var} \hat{\tilde{a}} \quad \frac{1-b_1}{b} \tilde{a} + \frac{b_1 - b}{b} a_0 + (\frac{M-m}{m(M-1)} F - \frac{1-b_1}{bN_1}) \sum_{j \geq 1} a_j^2 - \frac{1-b}{bN_0} a_0^2 .$$

Analogt til betraktningene på stratumnivå innføres vektoren

$$\underset{\sim}{\mu} = (\mu_0, \mu_1, \ldots),$$

der

$$\mu_i = a_i / \tilde{a}, \qquad i = 0, 1, \ldots .$$

Vi definerer den sekundære homogenitetsparameter ved

$$H = \sum_{j \geq 0} \mu_j^2 .$$

Tilsvarende til den primære homogenitetsparameter, er H et mål for variasjonene mellom strataene. Variasjonsområdet for μ blir

$$B = \{\underset{\sim}{\mu} \mid \Sigma \mu_i = 1, \quad 0 \leq \mu_i \leq \frac{N_i}{\tilde{a}}\}$$

fordi $a_i \leq N_i$. Vi skal se spesielt på situasjonen hvor $N_1 > \tilde{a}$ slik at variasjonsområdet blir

$$B' = \{\underset{\sim}{\mu} \mid \Sigma \mu_i = 1, \ 0 \leq \mu_i\}.$$

Med $\underset{\sim}{\mu}$ og H innsatt får (3.3) formen

$$\text{(3.4)} \qquad \operatorname{Var} \hat{\tilde{a}} \leq \frac{1-b_1}{b} \tilde{a} + \frac{b_1-b}{b} \mu_0 \tilde{a} + \left(\frac{M-m}{m(M-1)} F - \frac{1-b_1}{bN_1}\right)(H-\mu_0^2)\, \tilde{a}^2 - \frac{1-b}{bN_0} \tilde{a}^2 \mu_0^2 .$$

Når F og H holdes konstant mens μ_0 varierer, vil

$$\mu_0' = \frac{(b_1-b)}{2b\left(\frac{M-m}{m(M-1)} F - \frac{1-b_1}{bN_1} + \frac{1-b}{bN_0}\right) \tilde{a}}$$

maksimere høyre siden av (3.4).

Ved innsetting av μ_0' blir leddet som avhenger av μ_0 lik

$$\frac{(M-m)^2}{4m^2\left(\frac{M-m}{m(M-1)} F - \frac{1-b_1}{bN_1} + \frac{1-b}{bN_0}\right)} .$$

Leddet

$$\frac{1-b_1}{bN_1} - \frac{1-b}{bN_0}$$

vil som regel være lite sammenliknet med

$$\frac{M-m}{m(M-1)} F,$$

slik at uttrykket ovenfor blir av størrelsesorden

$$\frac{M - m}{4\, m\, F} M \leq \frac{40}{F}. \qquad (m = 2)$$

Vi vil anta at $F > 0.1$, $\tilde{a} > 1\,000$ og $b \leq 1/400$ slik at sammenliknet med

$$\frac{1 - b_1}{b} \tilde{a}$$

blir leddet

$$\frac{(M - m)M}{4\, m\, F}$$

neglisjerbart.

Med god tilnærmelse kan vi derfor sette

$$\text{(3.5)} \quad \text{Var}\, \hat{\tilde{a}} \leq \frac{1-b_1}{b} \tilde{a} + \{\frac{M-m}{m(M-1)} F - \frac{1-b_1}{b\, N_1}\} H \tilde{a}^2.$$

4. Ikke stratifisert utvalg

Vi vil se på situasjonen hvor stratifiseringen er uten betydning for de kjennetegn som undersøkes. Siden stratifisering ikke gir noe dårligere resultat (dvs. større varians) enn ikke-stratifisering (med Byråets utvalgsplan), vil vi basere våre betraktninger på variansuttrykket for ikke-stratifisert utvalg. Poenget ved å behandle denne situasjonen spesielt, er at vi er i stand til å redusere antall homogenitetsparametre til én mens vi i avsnittet foran opererte med to.

Vi har altså totrinns utvalg med et antall primærområder som vi kaller $L + 1$, der $L = 43$. $M = 1\,462$. Fra disse trekkes $\ell + 1$ områder rent lotterisk, med $\ell = 43\, m$. L av primærområdene er like store, mens ett "primærområde" i Oslo, er ca. 7 M ganger så stort som et "vanlig" primærområde.

La nå

$$a(j,k) = \begin{cases} 1 \text{ dersom den k-te IO i primærområde j har et spesielt kjennetegn,} \\ 0 \text{ ellers.} \end{cases}$$

$$a(j) = \sum_k a(j,k), \qquad a_0 = a(0), \quad N(0) = N_0,$$

$$a = \sum_{j=1}^{L} a(j), \qquad \tilde{a} = a + a_0,$$

$$\bar{a}(j) = a(j)/N(j), \qquad \bar{a} = a/L.$$

Vi observerer at annentrinns utvalgsbrøk er

$$\frac{L}{\ell}\, b = \frac{M}{m}\, b = b_1 \qquad \text{(Byråets utvalgsplan)},$$

slik at ved å benytte (2.1) får vi

$$\text{Var}\ \hat{\bar{a}} = \frac{1-b_1}{b} \sum_{j\geq 1} N(j)\, \bar{a}(j)\{1 - \bar{a}(j)\} + \frac{L-\ell}{b(L-1)}\, b_1 \sum_{j\geq 1} \{a(j) - \bar{a}\}^2$$

Siden $N(j) = N(1) = \bar{N}_1$ for $j > 0$,

finnes den totale varians av uttrykket ovenfor og (3.1):

$$(4.1)\quad \text{Var}\ \hat{\tilde{a}} = \frac{1-b_1}{b}\, a - \frac{L-\ell}{bL(L-1)}\, b_1 a^2 + \left(\frac{L-\ell}{b(L-1)}\, b_1 - \frac{1-b_1}{bN(1)}\right) \sum_{j>0} a(j)^2 + \frac{1-b}{b}\, a_0\left(1 - \frac{a_0}{N_0}\right).$$

Vi innfører den <u>totale homogenitetsparameter</u>

$$R = \sum_{j\geq 0} \gamma_j^{\,2},$$

der

$$\gamma_j = \frac{a(j)}{\tilde{a}}, \qquad j = 0, 1, \ldots;$$

og

$$\chi = (\gamma_0, \gamma_1, \ldots),$$

med variasjonsområde

$$C = \left\{ \chi \;\middle|\; \sum_{j\geq 0} \gamma_j = 1, \quad 0 \leq \gamma_j \leq \frac{N(j)}{\tilde{a}} \right\}.$$

Med χ og R innsatt får vi ved å bruke $(L-1)/L \approx 1$;

$$(4.2)\quad \text{Var}\ \hat{\tilde{a}} = \frac{1-b_1}{b}\, \tilde{a} + \frac{b_1-b}{b}\, \tilde{a}\, \gamma_0 + \left(\frac{L-\ell}{\ell} - \frac{1-b_1}{bN(1)}\right)(R-\gamma_0^{\,2})\, \tilde{a}^2 - \frac{L-\ell}{\ell(L-1)}\,(1-\gamma_0)^2\, \tilde{a}^2 - \frac{1-b}{bN_0}\, \gamma_0^{\,2}\, \tilde{a}^2.$$

Tilsvarende maksimeringen av (3.4) m.h.p. γ_0 finner vi at (4.2) har sitt maksimum for

$$\gamma_0^{*} = \frac{\frac{1}{L-1} - \frac{1}{2\tilde{a}}}{1 + \frac{1}{L-1} - \frac{\ell}{L-\ell}\left(\frac{1-b_1}{bN(1)} - \frac{1-b}{bN_0}\right)} \approx \frac{1}{L},$$

altså,

$$\operatorname{Var}\hat{\tilde{a}} \leq \frac{1-b_1}{b}\tilde{a}\left(1-\frac{\tilde{a}}{LN(1)}\right) + \left(\frac{L-\ell}{\ell L} - \frac{1-b_1}{bN(1)L}\right)(LR-1)\tilde{a}^2$$

$$+ \frac{b_1-b}{bL}\tilde{a} + \frac{1}{L^2}\left\{\frac{L-\ell}{\ell} + \frac{1-b_1}{bN(1)} - \frac{1-b}{bN_0}\right\}\tilde{a}^2.$$

De siste to leddene er små sammenliknet med de andre leddene slik at vi tilnærmet har

$$(4.3)\quad \operatorname{Var}\hat{\tilde{a}} \leq \frac{1-b_1}{b}\tilde{a}\left(1-\frac{\tilde{a}}{LN(1)}\right) + \left(\frac{L-\ell}{\ell L} - \frac{1-b_1}{bN(1)L}\right)(LR-1)\tilde{a}^2.$$

5. Ekstreme variansverdier. "Design-effekten"

Det kan være interessant å belyse nærmere hvordan variansfunksjonen ser ut for ekstreme verdier av de parametre som inngår. La oss som eksempel studere var $\hat{a}$ som funksjon av $R^*(\chi) = \sum_{j=1}^{L} \gamma_j^2$.

R^* kan betraktes som en generalisert paraboloide i R^L. Den er symetrisk omkring "R-aksen" og er strengt voksende når $|\chi|$øker. Følgelig har $R^*(\chi)$ minimum og maksimum i de punkter som ligger henholdsvis nærmest og lengst borte fra origo.

La

$$C_0 = \{\, \chi \in C \mid \gamma_0 = 0\}.$$

Det punktet i C_0 som ligger nærmest origo er

$$\chi^* = (0, \frac{1}{L+1}, \frac{1}{L+1}, \ldots)$$

noe som gir

$$R^*(\chi^*) = \frac{L}{(L+1)^2} \approx \frac{1}{L+1} = \min R^*(\chi) \quad \chi \in C_0$$

La $k = [a/N(1)]$, der $[x]$ er heltallsverdien til x. Det betyr at k av γ_j-ene kan anta verdien $N(1)/a$. De punktene i C_0 som ligger lengst fra origo er punktene hvor koordinatene er permutasjoner av koordinatene i vektoren

$$\chi^{**} = (0, \underbrace{\frac{N(1)}{a}, \frac{N(1)}{a}, \ldots \frac{N(1)}{a}}_{\text{k-komponenter}}, 1 - \frac{kN(1)}{a}, 0, \ldots, 0)$$

La $\delta = a/N(1) - k$. Da er $0 \leq \delta < 1$, og vi får

$$(5.5) \quad \max_{C_0} R^x(\chi) = \frac{N(1)}{a} - \delta(1-\delta)\left\{\frac{N(1)}{a}\right\}^2 .$$

Når $\chi \in C_0$ er $\tilde{a} = a$. Ulikheten (4.3) medfører derfor

$$(5.6) \quad \mathrm{Var}\,\hat{a} \leq \frac{1-b_1}{b}\, a\,(1-\frac{a}{L})$$

når $\chi = \chi^x$, og

$$(5.7) \quad \mathrm{Var}\,\hat{a} \leq \frac{L-\ell}{\ell L}\{N(1)\,L\,a - a^2\} - N(1)^2\left\{\frac{L-\ell}{\ell} - \frac{1-b_1}{bN(1)}\right\}\delta(1-\delta)$$

når $\chi = \chi^{xx}$. La

$$p = a/N',$$

der $N' = N - N_0 = N(1)L$. Siden $\chi \in C_0$, får vi

$$\mathrm{Var}\,\hat{a} \leq \frac{L-\ell}{\ell} N(1)^2 Lp(1-p) - N(1)^2\left\{\frac{L-\ell}{\ell} - \frac{1-b_1}{bN(1)}\right\}\delta(1-\delta).$$

Når $p(1-p)$ ikke er for liten, ser vi at det siste leddet i høyre side ovenfor blir mye mindre enn det første. Vi har derfor tilnærmet

$$(5.8) \quad \mathrm{Var}\,\hat{a} \leq \frac{L-\ell}{\ell L} N'^2\, p(1-p).$$

Byråets praksis er som nevnt å benytte variansformelen

$$(5.9) \quad \mathrm{Var}\,\hat{a} \approx 1.5\,\frac{N'^2}{n'}\,p(1-p)$$

der $n' = n-n_0$. Altså "tilsvarer" faktoren $\frac{L-\ell}{\ell L}$ i (5.8) faktoren $\frac{1.5}{n'}$ i (5.9). Med $m = 6$, blir disse faktorene like når

$$n' = 1.5\,\frac{L\ell}{L-\ell} = 470.$$

Vi får videre av (5.8) og (5.9) at følgende ulikhet gjelder for "Design"-effekten

$$(5.10) \quad n'\,\frac{\mathrm{Var}\,\hat{a}}{N'^2\,p(1-p)} \leq \frac{L-\ell}{\ell L}\,n' = \frac{n'}{313}\,.$$

Følgende tabell viser størrelsen av dette forholdet når n' varierer.

n'	n'/313
470	1.5
1 500	4.8
2 000	6.4
3 000	9.6
5 000	16

Vi ser at når n' øker, kan vi i verste fall ha en "design"-effekt som langt overstiger 1.5. Imidlertid er en slik fordeling av a-verdiene som antatt her, så spesiell at den vel må sies bare å ha teoretisk interesse.

6. Nedre grense for estimand

Når $\hat{\tilde{a}}$ forutsettes tilnærmet normalfordelt, vil en ha

$$\tilde{a} - t\sqrt{\operatorname{Var}\hat{\tilde{a}}} \leq \hat{\tilde{a}} \leq \tilde{a} + t\sqrt{\operatorname{Var}\hat{\tilde{a}}}$$

med sannsynlighet $1-\varepsilon$ når t er $(1 - \frac{\varepsilon}{2})$-fraktilen i den standardiserte normalfordelingen. Analogt til Hoem [1] gir (3.5) at den relative andel som avviket i grensene ovenfor utgjør av $\tilde{a}$ (konfidenskoeffisienten) er

$$(6.1) \qquad g(\tilde{a}) = \frac{t\sqrt{\operatorname{Var}\hat{\tilde{a}}}}{\tilde{a}} \leq t\left\{\frac{1-b_1}{b\,\tilde{a}} + \left[\frac{M-m}{m(M-1)}F - \frac{1-b_1}{b\,N_1}\right]H\right\}^{\frac{1}{2}}.$$

Dersom det forlanges at

$$g(\tilde{a}) \leq \theta,$$

må

$$(6.2) \qquad \tilde{a} \geq \frac{1-b_1}{b\left\{\left(\frac{\theta}{t}\right)^2 - \left[\frac{M-m}{m(M-1)}F - \frac{1-b_1}{b\,N_1}\right]H\right\}} = f(F,H).$$

Dersom vi bare er interessert i a-verdien for et stratum, f.eks. stratum nr. r, får vi av (2.4) at

$$a_r \geq \frac{1-b_r}{b\left\{\left(\frac{\theta}{t}\right)^2 + \frac{1-b_r}{N_r} - \left[\frac{M_r-m_r}{m_r(M_r-1)} - \frac{1-b_r}{N_r b}\right](M_r K_r - 1)\right\}}.$$

Siden $(1-b_r)/N_r$ er mye mindre enn $(\frac{\theta}{t})^2$, kan den siste ulikheten forenkles til

$$(6.3)\quad a_r \geqq \frac{1 - b_r}{b\{(\frac{\theta}{t})^2 - \left[\frac{M_r - m_r}{m_r(M_r - 1)} - \frac{1-b_r}{N_r b}\right](M_r K_r - 1)\}}.$$

Av (4.3) får vi for ikke-stratifisert utvalg når tilnærmelsen

$$(\frac{\theta}{t})^2 + \frac{1-b_1}{LN(1)} \approx (\frac{\theta}{t})^2;$$

benyttes;

$$(6.4)\quad \tilde{a} \geqq \frac{1-b_1}{b\{(\frac{\theta}{t})^2 - \left[\frac{L-\ell}{\ell L} - \frac{1-b_1}{bN(1)L}\right](LR-1)\}}.$$

Vi har dermed funnet en formel som gir oss en nedre grense for estimand som funksjon av homogenitetsparametrene. Problemet er nå den a priori vurderingen av disse parametrene. I denne forbindelse kan det være hensiktsmessig å bruke en annen parameter enn ρ_{ij}. Dersom vi i steden definerer ℓ_{ij} ved

$$(6.5)\quad a_i(j) = \bar{a}_i + \ell_{ij}\,\tilde{a}_i = \frac{1+\ell_{ij}}{M_i}\,a_i,$$

blir

$$\rho_{ij} = \frac{1+\ell_{ij}}{M_i}.$$

Vektoren $\underset{\sim}{\ell}_i = (\ell_{i1}, \ell_{i2}, \ldots, \ell_{iM})$

får variasjonsområde

$$D = \{\underset{\sim}{\ell}_i \mid \sum_j \ell_{ij} = 0,\ -1 \leqq \ell_{ij} \leqq \min(\frac{N_i}{a_i} - 1, M_i - 1)\}.$$

Dessuten blir

$$(6.6)\quad MK_i - 1 = \frac{1}{M}\sum_j \ell_{ij}^2.$$

La ε_1, ε_2 være slik at for q primærområder i hvert stratum (unntatt BT) er

$$|\ell_{ij}| \leqq \varepsilon_1 \qquad (1 \leqq i \leqq 40)$$

og for M - q primærområde i hvert stratum er

$$|\alpha_{ij}| \leq \varepsilon_2. \qquad (1 \leq i \leq 40)$$

Det vil si at for q primærområder i hvert stratum varierer a-verdien høyst 100 ε_1 % omkring den gjennomsnittlige a-verdi i stratumet, og i de resterende M - q primærområder varierer a-verdien høyst 100 ε_2 % omkring gjennomsnittet. Av denne antagelsen følger at $MK_i - 1$ blir begrenset av

$$(6.7) \quad F = \varepsilon_2^2 + \frac{q}{M}(\varepsilon_1^2 - \varepsilon_2^2) \qquad (1 \leq i \leq 40).$$

La oss se nærmere på ulikheten (3.2). Vi vil benytte tilnærmelsene

$$\frac{M-m}{m(M-1)} - \frac{1-b_1}{N_1 b} \approx \frac{M-m}{m\,M}, \quad \frac{1}{M-2} \approx \frac{1}{M-1} \approx \frac{1}{M}.$$

Etter multiplikasjon med m M/(M-m) blir siste del av (3.2)

$$(6.8) \quad MK_i - 2 \leq \frac{1-b_1}{b\,N_1} \cdot \frac{m\,M}{M-m} + F, \qquad 41 \leq i \leq 46.$$

Når m = 6 og b = 1/250, blir

$$\frac{1-b_1}{b\,N_1} \cdot \frac{m\,M}{M-m} \approx 0{,}013.$$

Vi kan derfor tilnærme kravet (3.2) med

$$(6.9) \quad \begin{cases} M K_i - 1 \leq F, & 1 \leq i \leq 40 \\ M K_i - 1 \leq 1 + F, & 41 \leq i \leq 46. \end{cases}$$

Vi skal ikke her gå nærmere inn på fastleggelsen av H, F og R, siden det ikke foreligger tilstrekkelig empirisk materiale på det nåværende tidspunkt. Det er derfor viktig at det blir gjort beregninger over størrelsen av disse parametrene. Jeg kan tenke meg to måter dette kan gjøres på:

i) En estimerer homogenitetsparametrene for hver intervjuundersøkelse på grunnlag av estimatene for $a_i(j)$, a_i og $\tilde{a}$.

ii) En benytter materialet fra tidligere undersøkelser til å gjøre et "stort antall" beregninger for å belyse hvordan homogenitetsparametrene varierer. På grunnlag av disse beregninger fastsettes et sett (F,H) (eventuelt flere) som brukes for alle intervjuundersøkelser (eventuelt for alle undersøkelser av samme type).

Til slutt tar vi med tabellen som viser sammenhengen mellom de ulike parametrene og nedre grense for estimanden $\tilde{a}$.

Ved bruk av "1,5 regelen" kom en fram til en nedre grense for $\tilde{a}$ lik 10 000. [1] (θ = 38 %). Vi ser av tabell 2 at dette tilsvarer følgende verdier av (F,H)

$\{(15\ \%, 1);\ (25\ \%, \frac{1}{2});\ (100\ \%, \frac{1}{8});\ (200\ \%, \frac{1}{16});\ (300\ \%, \frac{1}{16})\}$.

Tabell 1. Verdier av F for varierende ε_1, ε_2 og q der F er gitt ved (6.7)

q=12	ε_1%	40	80	100	150	200	50	80	100	150	200	250	300	70	100	150	200
	ε_2%	40	40	40	40	40	50	50	50	50	50	50	50	70	70	70	70
	F %	16	33	45	90	150	25	40	50	86	156	235	331	49	67	111	172
q=6	ε_1%	60	100	150	200	250	50	70	100	150	200	250	300	350	400	450	500
	ε_2%	40	40	40	40	40	50	50	50	50	50	50	50	50	50	50	50
	F %	20	31	54	85	126	25	21	39	61	93	133	183	240	309	385	421
q=6	ε_1%	60	100	150	200	250	300	350	100	150	200	250	300	350	400	450	500
	ε_2%	30	30	30	30	30	30	30	70	70	70	70	70	70	70	70	70
	F %	19	41	85	146	225	321	435	58	81	112	153	202	261	328	405	490
q=8	ε_1%	70	100	150	200	250	300	70	100	150	200	250	300	350	400	450	500
	ε_2%	40	40	40	40	40	40	50	50	50	50	50	50	50	50	50	50
	F %	24	36	66	108	162	228	31	43	73	115	169	235	313	403	505	619
q=8	ε_1%	100	150	200	250	300	350	400	450	500							
	ε_2%	70	70	70	70	70	70	70	70	70							
	F %	61	91	133	187	253	331	421	523	637							

Tabell 2. Verdier av f(F,H), gitt ved likning (6.2), for varierende F og H når θ = 40 %, m = 6, M = 34, N_1 = 2 000 M, b = 1/250 og nivået ε = 5 %.

H	1	1	1	1/2	1/2	1/2	1/2	1/2	1/2	1/4	1/4
F%	15	20	25	15	20	25	30	35	40	25	40
f	10 000	14 700	25 700	7 500	8 400	9 600	11 000	13 300	16 500	7 300	8 700
H	1/4	1/4	1/4	1/8	1/8	1/8	1/8	1/8	1/8	1/16	1/16
F%	50	60	70	25	50	80	100	120	150	50	100
f	10 000	11 600	14 000	6 500	7 400	8 800	10 000	11 800	16 000	6 600	7 500
H	1/16	1/16	1/16	1/25	1/25	1/25	1/25	1/25			
F%	150	200	250	200	250	300	350	400			
f	8 600	10 200	12 600	8 000	8 800	9 800	11 000	12 600			

Referanser:

[1] Jan M. Hoem, (1971). "Hvor små befolkningsgrupper kan vi gi rimelig sikre tall for i tabellene fra arbeidskraft-undersøkelsene?" Maskinskrevet notat JMH/GH, 7/12-71.

[2] Steinar Tamsfoss, (1970). "Om bruk av stikkprøver ved kontoret for intervjuundersøkelser." Statistisk Sentralbyrå, Artikkel 37.

JMH/GH, 4/5-72

A p p e n d i k s

Lett bearbeidet utdrag fra:

VARIANSBEREGNINGER VED INTERVJUUNDERSØKELSER

IV. FASTLEGGELSE AV TOTAL UTVALGSBRØK b

Av Jan M. Hoem

1. Den "klassiske" teorien

Ett av trekkene ved den vanlige beskrivelsen av særegenhetene ved Byråets normale utvalgsplan er at antall intervjuenheter sies å være bestemt ved

$$n_{ir} = b_i N_i(j_{ir}), \tag{A.1.1}$$

der

$$b_i = b\, M_i/m_i, \tag{A.1.2}$$

med b som en "total utvalgsbrøk" som fastlegges stokastisk uavhengig av utvalgstrekkingen. Samlet antall utvalgsenheter blir etter dette

$$n = \sum_{i,r} n_{ir}(J_{\sim i}) = b \sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir}). \tag{A.1.3}$$

n blir da en stokastisk variabel. Dens forventningsverdi er

$$En = bN,$$

der N er det samlede antall intervjuenheter i landet. Utvalgsbrøken b har altså tolkningen

$$b = En/N.$$

2. Byrået bruker egentlig en annen fremgangsmåte

Den metoden vi har beskrevet ovenfor, innebærer at man egentlig ikke har skikkelig kontroll med n. Denne størrelsen kan i praksis avvike ganske mye fra sin forventningsverdi bN. Eksempelvis kan det hende at man tar sikte

på en utvalgsstørrelse på 3 000 (dvs. man fastlegger b = 3 000 / N), mens prosedyren ovenfor resulterer i at n er av størrelsesorden 3 500. Dette vil ikke være akseptabelt for Intervjukontoret, som i et slikt tilfelle vil redusere utvalgsstørrelsen fra det "opprinnelige" ca. 3 500 til det forønskede ca. 3 000.

Intervjukontoret følger altså egentlig ikke det opplegg som er beskrevet bl.a. i Byråets Artikkel nr. 37 ("Om bruk av stikkprøver ved Kontoret for intervjuundersøkelser, Statistisk Sentralbyrå"). Det vanlige formelverk kan derfor ikke uten videre benyttes ved Byråets intervjuundersøkelser.

Hvis man primært ønsker å ha kontroll over n, kan det være rasjonelt å fastlegge n stokastisk uavhengig av utvalgstrekkingen (f.eks. fiksere n til 3 000), og så bestemme b slik at (A.1.3) fortsatt gjelder. Mens (A.1.3) opprinnelig var en relasjon som definerte n, går man da over til å ta n som gitt og bruke (A.1.3) til å definere b, som da blir

$$(A.2.1) \qquad b = n \Big/ \sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir}).$$

Dette betyr at b nå blir en stokastisk variabel, avhengig av alle J_{ir}. Hvis n_{ir} fortsatt bestemmes ved (A.1.1) og (A.1.2), blir også denne en stokastisk variabel, avhengig av alle J_{ir}.

Men da blir utvalgene fra de ulike strataene stokastisk avhengige, i strid med forutsetningene for det vanlige formelverket.